CAN VERIFICATION SUSTAIN SKILL?

Anonymous authors
Paper under double-blind review

ABSTRACT

A common way to make AI delegation reliable is to have AI generate an output that a human verifies and corrects. A static view of this workflow treats verification competence as fixed. Over time, however, practice and non-use can change this competence. We model a worker whose skill determines manual and verification reliabilities and thereby shapes workflow-specific performance losses and completion costs. We study how skill evolves when the worker chooses among manual work, verified delegation, and pure delegation to maximize current utility by trading off performance loss against completion cost. Current skill determines workflow choice, which shapes future skill. Under verified delegation, reviewing provides verification practice, whereas generative practice occurs only when the worker detects and repairs an AI error. We characterize the global dynamics and identify three verification-related regimes: collapse to zero skill, an interior skill level sustained by verified delegation, and competence sustained at a manual–verified switching boundary. Crucially, verified delegation can remain optimal over a skill range while no positive skill level is sustainable. Adding verified delegation can improve current utility yet eliminate an otherwise high-skill basin. Holding learning dynamics fixed, workflow incentives alone can produce all three regimes. Better AI can further reduce retained human skill by reducing manual practice and the errors that trigger repair-based learning. The regimes persist over nontrivial learning-parameter ranges and under additional learning from correct AI outputs, while task-level randomness can drive trajectories near a stable verified equilibrium into pure delegation. Thus, keeping a human in the loop need not preserve the capability required for effective future oversight.

1 INTRODUCTION

A common response to the risks of AI delegation is to keep a human in the loop. AI may generate a draft or a solution, while a qualified worker remains responsible for checking and correcting it. Such workflows promise to combine the productivity gains of delegation with the reliability of human oversight, particularly when workers are better at reviewing AI outputs than producing the underlying work independently (Huang et al., 2026). Yet effective oversight cannot be taken for granted: scrutiny of AI outputs can weaken under cognitive load, task complexity, and time pressure (Gaube et al., 2021; Bergman, 2025), while related work documents over-reliance and automation-related failures (Parasuraman et al., 2000; Lyell & Coiera, 2017; Passi & Vorvoreanu, 2022).

There is a further problem that is inherently dynamic. In human-in-the-loop workflows, effective verification depends on human competence. Detecting an omitted case, recognizing a faulty argument, checking a computation, or repairing an incorrect output can draw on cognitive processes closely related to those used in producing the work itself (Campbell, 1987; Zbrodoff & Logan, 1990; Campbell & Tarling, 1996). Durable competence depends on continued retrieval and practice (Roediger & Karpicke, 2006; Karpicke & Roediger, 2008; Cepeda et al., 2006). Delegating generation may therefore preserve a worker's formal role as a reviewer while reducing the practice through which effective human review remains possible. This concern is consistent with evidence that cognitive offloading can improve immediate performance while diminishing memory (Risko & Gilbert, 2016; Grinschgl et al., 2021), and with recent studies showing that generative-AI assistance can weaken learning or subsequent independent performance (Bastani et al., 2025; Kosmyna et al., 2025; Liu et al., 2026; Shen & Tamkin, 2026). These observations raise a basic question that we study in this paper: *Can human verification remain dynamically sustainable when the ability required for verification itself is endogenous?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Related work. Two closely related frameworks isolate complementary parts of the problem studied here. The composer–reviewer model of Huang et al. (2026) distinguishes manual work, verified delegation, and pure delegation, but holds human production and verification abilities fixed. The dynamic model of Huang & Vishnoi (2026), by contrast, allows human skill and AI reliance to co-evolve over repeated tasks, but treats delegation as a single action and therefore does not distinguish AI generation with human verification from pure delegation. The first model shows why verification can make delegation valuable; the second shows how repeated AI reliance can erode human skill. Neither studies whether verified delegation remains sustainable when the competence required for effective verification itself evolves with repeated workflow choice. A broader literature studies how AI use interacts with human learning, reliance, and cognitive offloading (Parasuraman et al., 2000; Risko & Gilbert, 2016; Grinschgl et al., 2021; Kosmyna et al., 2025; Liu et al., 2026; Shen & Tamkin, 2026). This work provides evidence that assistance and offloading can improve current performance while altering subsequent human capability, but does not jointly model repeated workflow choice, the distinction between generation and verification, and the evolution of the competence required for future verification. A fuller comparison with these literatures appears in Appendix A.

Our contributions. We develop a dynamic model of AI-assisted work in which human competence and workflow choice jointly evolve. A worker chooses among *manual work*, *verified delegation*, and *pure delegation* to maximize per-task utility (Eq. (1)). Current skill determines manual and verification reliabilities and workflow choice; the selected workflow then determines the learning signal. The learning mechanism separates generation from verification: manual work provides generative practice, pure delegation permits non-use decay, and verified delegation provides verification practice, with generative practice only when an AI error is detected and repaired (Eq. (5)).

We ask whether verification sustains the skill needed for continued human oversight. The main difficulty is that learning changes workflow incentives: a worker may leave a workflow before reaching its stable learning equilibrium. Theorem 3.1 characterizes these choices: pure delegation occupies a single lower-skill interval, while verification can be optimal in disconnected regions, allowing AI to compensate for weak production at lower skill and complement reliable verification at higher skill. Building on this partition, our main result characterizes global skill convergence and its basins by identifying active branch equilibria and additional attractors created by workflow switching (Theorem 3.2). The theorem reveals how verification introduces new paths to collapse, stagnation, and boundary-supported competence. Proposition D.1 shows that changing only workflow incentives can select among these three regimes from the same initial skill, with all three learning drifts fixed.

In the collapse regime, verification can remain optimal over a high-skill region even though skill converges to zero from every interior initial condition (Corollary 3.3). Even a highly skilled worker can therefore begin losing skill while inspecting and repairing AI outputs, eventually switch to pure delegation, and collapse through continued non-use. Within verified delegation, verification practice and error-triggered repair can balance non-use decay, sustaining an interior equilibrium below full skill. At a manual–verified boundary, manual practice can offset skill loss under verification, creating an attractor even though neither adjacent branch has an equilibrium there. See Figure 1 for a parametric illustration. Holding all primitives fixed, adding verified delegation can improve current utility while eliminating an otherwise high-skill basin (Proposition D.2). These results suggest that workers should consider learning alongside current performance and effort, and retain independent practice when verification and repair alone cannot maintain skill.

We next study how better AI affects retained human skill. Both workflow choice and learning change with AI quality, so fixed-workflow comparisons are insufficient. Theorem 3.4 identifies two channels: better AI expands pure delegation and contracts manual work, while fewer errors reduce repair practice under verification. A stable verified equilibrium is therefore nonincreasing in AI quality while active. Moreover, better AI cannot increase long-run skill from the same initial state.

Section 4 examines sensitivity to learning parameters and learning from correct reviewed outputs. Task-level simulations show that mean stability need not imply finite-task durability: random repair opportunities can trigger escape from a stable verified equilibrium into pure delegation. An illustrative data-to-parameter mapping using Bastani et al. (2025) shows how workflow observations can inform selected model inputs.

Taken together, the results show that keeping a human in the loop can improve current outcomes without preserving the skill needed for effective future oversight.

2 MODEL

We study repeated AI-assisted work over comparable tasks. Let $\mathcal{A} := \{M, V, P\}$ denote the workflows: under manual work M, the worker generates and self-checks the output; under verified delegation V, AI generates a draft that the worker reviews and repairs if an error is detected; and under pure delegation P, the AI output is accepted without review. On each task the worker chooses a workflow to maximize utility, and experience and feedback shape future skill. Appendix B.1 presents the notation in Table 1 and illustrates the workflows in Figure 3.

Worker and AI skill. Let $\theta \in [0, 1]$ denote latent task-specific worker skill, normalized so that larger θ corresponds to greater competence. This scalar state captures the task knowledge underlying both manual production and verification. The corresponding reliability functions, $r_m(\theta)$ and $r_v(\theta)$, defined below, allow the same latent skill to translate into different levels of production and verification ability.

This latent-trait representation is standard in models of ability (Embretson & Reise, 2000) and is mathematically tractable. Let $\theta_a \in [0, 1]$ denote task-specific AI skill, held fixed over a learning horizon of comparable tasks (Dell'Acqua et al., 2023).

Utility-based workflow. The worker uses calibrated beliefs to trade off expected final loss against completion cost. Let $\ell_A(\theta; \theta_a)$ and $\mathrm{cost}_A(\theta; \theta_a)$ denote these quantities under $A \in \mathcal{A}$, and let $\gamma \geq 0$ convert cost into loss units (Guo et al., 2024). The per-task utility and optimal workflow are

$$U_A(\theta; \theta_a) = -\ell_A(\theta; \theta_a) - \gamma\, \mathrm{cost}_A(\theta; \theta_a), \qquad A^*(\theta; \theta_a) \in \arg\max_{A \in \mathcal{A}} U_A(\theta; \theta_a), \tag{1}$$

with ties broken by an arbitrary but fixed rule.[1] We assume calibrated beliefs and myopic workflow choice with respect to skill formation: future learning does not enter (1).

Reliability and performance loss. Under pure delegation P, let $\varepsilon_a(\theta_a) \in [0, 1]$ denote the probability of a loss-relevant AI error, and let $\bar{\ell}_a(\theta_a) \in [0, 1]$ denote its conditional mean severity. We normalize the loss of a correct AI output to zero. Thus, the expected loss under P is

$$\ell_P(\theta_a) := \varepsilon_a(\theta_a)\bar{\ell}_a(\theta_a). \tag{2}$$

We assume that ε_a and $\bar{\ell}_a$ are continuously differentiable and nonincreasing in θ_a. Error frequency, rather than conditional severity, determines how often verification can lead to detection and repair. Under manual work M, let $r_m : [0, 1] \to [0, 1]$ denote manual-output reliability, the normalized expected quality of the worker's final output after ordinary self-checking. We define the corresponding expected loss as

$$\ell_M(\theta) := 1 - r_m(\theta).$$

Under binary loss, $r_m(\theta)$ and $\ell_M(\theta)$ are the probabilities of success and failure, respectively. We assume that r_m is continuously differentiable, nondecreasing, and concave in θ. Higher skill yields more reliable manual output, while concavity captures diminishing gains as performance approaches its ceiling, consistent with power and exponential laws of practice (Newell & Rosenbloom, 1981; Heathcote et al., 2000). A convenient family is $r_m(\theta) = 1-(1-\rho\theta)^{p_m}$, with $\rho \in (0, 1]$ and $p_m \geq 1$. Here ρ captures task tractability, and p_m represents an effective number of opportunities to produce or correct the output.

Under verified delegation V, let $r_v : [0, 1] \to [0, 1]$ denote *verification reliability*: the probability that the worker detects a loss-relevant AI error, conditional on such an error occurring. For tractability, we assume a separable detection technology, as in Huang et al. (2026), so that conditional detection depends only on worker skill. This abstracts from task-level heterogeneity in verification difficulty (Lyell & Coiera, 2017). We also rule out false positives in the model. The probability that an AI error occurs and is detected is therefore $q(\theta; \theta_a) := \varepsilon_a(\theta_a)r_v(\theta)$. If an error is missed, the submitted output retains its conditional mean loss $\bar{\ell}_a(\theta_a)$. If it is detected, we model repair as manual corrective production, so the repaired output has mean loss $\ell_M(\theta)$. Hence,

$$\ell_V(\theta; \theta_a) = \varepsilon_a(\theta_a)\big[(1 - r_v(\theta))\bar{\ell}_a(\theta_a) + r_v(\theta)\ell_M(\theta)\big]. \tag{3}$$

For learning, we separately model each repair as providing the same generative-learning signal as manual production. Appendix B.2 shows that (3) continues to hold under a weaker conditional-mean restriction.

[1] At workflow boundaries, we use the solution convention specified in Appendix C.2.

We assume that r_v is continuously differentiable, nondecreasing, and concave in θ: higher skill improves error detection, with diminishing gains as reliability approaches its ceiling. A concrete example is $r_v(\theta) = 1 - (1 - \rho\theta)^{p_v}$, $p_v \in \mathbb{N}_{\geq 1}$, which gives the probability of detection in at least one of p_v conditionally independent inspection opportunities, each with success probability $\rho\theta$ (Koopman, 1956; Porter & Votta, 1998). Allowing p_m and p_v to differ lets the same underlying skill map differently into manual-output reliability r_m and verification reliability r_v.

Completion cost. Let $k_a \geq 0$ denote the per-task AI-use cost, and let $c_g, c_s, c_v : [0,1] \to \mathbb{R}_{\geq 0}$ denote the costs of generation from scratch, ordinary self-checking, and inspecting an AI-generated output, respectively. Under verified delegation, a detected error requires repair; let $\lambda \in [0,1]$ scale the effort of repair relative to generation from scratch. Since detection occurs with probability $q(\theta; \theta_a)$, the workflow-specific expected costs are

$$\mathrm{cost}_M = c_g(\theta) + c_s(\theta), \qquad \mathrm{cost}_V = k_a + c_v(\theta) + \lambda q(\theta;\theta_a) c_g(\theta), \qquad \mathrm{cost}_P = k_a. \tag{4}$$

Here, c_v includes inspection, while λ discounts repair effort only; repair quality and the resulting learning signal are unchanged. We assume that c_g, c_s, c_v are continuously differentiable and nonincreasing in skill, consistent with evidence that task-completion time falls and eventually flattens with practice (Newell & Rosenbloom, 1981; Heathcote et al., 2000). A convenient family is $c_j(\theta) = k_j + b_j(1-\theta)^d$, $j \in \{g, s, v\}$, with $k_j, b_j \geq 0$ and $d \geq 1$. Here k_j is the irreducible effort floor, b_j scales the skill-sensitive component, and d controls how quickly cost approaches that floor as skill rises. The parametric illustration uses these families for r_m, r_v, c_j; see Figure 1 and Appendix D. These functional forms are illustrative; the general results below use the stated shape restrictions rather than the particular parametric families.

Workflow-dependent skill dynamics. Let $\theta = \theta(t)$, with time normalized by task opportunities and $\theta(0) = \theta_0$. Skill evolves through three mechanisms. We model *generative learning* from production by $-\ell'_M(\theta) \geq 0$, motivated by error-driven learning and the generation effect (Rescorla & Wagner, 1972; Slamecka & Graf, 1978; Bjork, 1994). We model *non-use* by the decay term $-\Delta\theta$, $\Delta > 0$, consistent with learning-and-forgetting models and evidence on independent performance after AI assistance (Pavlik & Anderson, 2005; Liu et al., 2026; Shen & Tamkin, 2026). Finally, we model *verification practice* by $\kappa(1-\theta)$, $\kappa \geq 0$, allowing active inspection and self-explanation to support learning without independent production (Chi et al., 1989).

The selected workflow determines exposure to these mechanisms. Under M, generative learning and verification practice through self-checking occur on every task; under P, only non-use occurs. Under V, verification practice occurs on every task, while generative learning follows a detected and repaired AI error with probability $q(\theta;\theta_a)$; otherwise generative skill is exposed to non-use.

Using the bounded natural-gradient scaling $\theta(1-\theta)$, which keeps both boundaries invariant, these workflow-specific learning forces yield

$$\dot\theta = f_{A^*}(\theta) = \begin{cases} -\theta(1-\theta)\big[\ell'_M(\theta) - \kappa(1-\theta)\big], & A^* = M, \\ -\theta(1-\theta)\big[q(\theta;\theta_a)\ell'_M(\theta) - \kappa(1-\theta) + (1-q(\theta;\theta_a))\Delta\theta\big], & A^* = V, \\ -\Delta\theta^2(1-\theta), & A^* = P. \end{cases} \tag{5}$$

Appendix B.3 derives (5) using the bounded natural-gradient construction (Amari, 1998; Huang & Vishnoi, 2026). We interpret the ODE as a continuous-time mean-field approximation to small per-task skill changes (Benaïm, 1999). Since $\ell'_M(\theta) \leq 0$, manual work cannot lower skill ($f_M(\theta) \geq 0$), while pure delegation cannot raise it ($f_P(\theta) \leq 0$). Under verified delegation, the sign of $f_V(\theta)$ depends on the balance between error-triggered learning, verification practice, and non-use decay. It follows from (5) that for any $\theta \in (0,1)$, $f_M(\theta) \geq f_V(\theta) \geq f_P(\theta)$. Utility maximization selects the active skill dynamics; skill does not follow the gradient of per-task utility.

Relation to prior models. Conditional on a fixed worker state, optimizing verification effort in Huang et al. (2026) reduces the delegation–verification problem to the three workflow categories M, V, P; here, an evolving skill state determines their losses and costs. Conversely, removing verified delegation and mixing the manual and pure-delegation branches recovers, under the corresponding normalization, the skill dynamics of Huang & Vishnoi (2026). That model closes the system through an adaptive continuous delegation level, whereas here current utility selects among discrete workflows. The new element is the verified-delegation branch: verification practice continues, while generative practice occurs only after a detected and repaired AI error, and skill dynamics can induce switching into or out of this branch. Appendices B.4–B.5 give the explicit reductions.

3 THEORETICAL RESULTS

This section characterizes the long-run worker skill induced by repeated utility-maximizing workflow choice. Unless otherwise stated, we fix the worker-reliability functions (r_m, r_v), the AI-error functions $(\varepsilon_a, \bar{\ell}_a)$, the activity-cost functions (c_g, c_s, c_v), and the parameters $(k_a, \lambda, \gamma, \kappa, \Delta)$.

We first fix θ_a and suppress it from the notation. For each workflow $A \in \mathcal{A}$, let $R_A := \{\theta \in (0,1) : U_A(\theta) > \max_{B \neq A} U_B(\theta)\}$ denote the strict action region of A. Away from workflow ties, the endogenous dynamics are $\dot{\theta} = f^\star(\theta) := f_{A^*}(\theta)$ for $\theta \in R_{A^*}$. Thus, current utility determines which branch is active, while the learning model determines the motion generated by that branch. A stable zero z of f_V is an active verified equilibrium only if $z \in R_V$, whereas an $M \to V$ boundary τ with $f_M(\tau) > 0 > f_V(\tau)$ can itself be attracting even though neither adjacent branch vanishes there. The relevant object is therefore the joint geometry of the action regions $\{R_M, R_V, R_P\}$ and the branch dynamics $\{f_M, f_V, f_P\}$.

Theorem 3.1 first characterizes the endogenous workflow partition. Our main result, Theorem 3.2, combines this partition with the branch dynamics to give a global characterization of limiting skill. Comparisons holding the learning drifts fixed isolate the effects of workflow incentives and the introduction of verified delegation (Propositions D.1 and D.2). Corollary 3.3 gives sufficient conditions for global collapse; the constructions show that these conditions can hold even while verified delegation remains optimal over a high-skill region. Finally, Theorem 3.4 characterizes the effects of AI improvement on workflow choice, learning within verified delegation, and long-run skill.

Characterization of worker action. We first determine the regions in which the three branch dynamics are active. Let $G(\theta) := \max\{U_M(\theta), U_V(\theta)\} - U_P(\theta)$ denote the utility gain from retaining human involvement rather than using pure delegation. The model primitives imply that $\{G > 0\}$ is an upper interval; see Lemma C.1. We therefore define the pure-delegation boundary by

$$\tau_P := \inf\{\theta \in (0,1) : G(\theta) > 0\}, \qquad \inf \varnothing := 1.$$

Thus, pure delegation P can occupy only the lower-skill portion of the state space.

Above τ_P, the remaining comparison is between manual work and verified delegation. Let $D(\theta) := U_V(\theta) - U_M(\theta)$. Unlike the comparison with pure delegation, the relative ranking of M and V can reverse as skill increases, so D need not be monotone. Denote the sign-changing zeros of D in $(\tau_P, 1)$ by $\tau^{(1)} < \cdots < \tau^{(n)}$. Set $\tau^{(0)} := \tau_P$, $\tau^{(n+1)} := 1$, and, when $\tau_P < 1$, let $\sigma \in \{-1, +1\}$ denote the sign of D at nonzero points immediately to the right of τ_P. We present the finite-n result below, which covers the workflow partition in Figure 1a; Appendix C.1 gives the general characterization.

Theorem 3.1 (Worker action; informal). *Pure delegation P is optimal for $\theta \in (0, \tau_P)$. For $\theta > \tau_P$ away from workflow ties,*

$$A^*(\theta) = \begin{cases} V, & \theta \in (\tau^{(j)}, \tau^{(j+1)}) \text{ and } \sigma(-1)^j > 0, \\ M, & \theta \in (\tau^{(j)}, \tau^{(j+1)}) \text{ and } \sigma(-1)^j < 0, \end{cases} \qquad j = 0, \ldots, n.$$

The threshold τ_P represents the end of pure delegation P, while $\tau^{(1)}, \ldots, \tau^{(n)}$ mark switches between manual work M and verified delegation V. The sign σ determines the first workflow above τ_P: the ordering is $P \to V \to M \to \cdots$ when $\sigma = 1$ and $P \to M \to V \to \cdots$ when $\sigma = -1$. This theorem permits verified delegation to occupy disconnected skill regions R_V. The same AI workflow can compensate for weak independent production at lower skill and become complementary to reliable detection and repair at higher skill. Separating production and verification therefore implies that AI use need not follow a single worker-skill cutoff.

The single P-boundary does not require $U_V - U_P$ to be globally monotone. Writing $U_V - U_P = qB - \gamma c_v$, where $B := \bar{\ell}_a - \ell_M - \gamma\lambda c_g$, once this gain is positive, $B > 0$; since q and B are nondecreasing in skill θ while c_v is nonincreasing in θ, it remains positive as θ further increases. Together with monotonicity of $U_M - U_P$, this rules out re-entry of P; see Appendix C.1.

Skill convergence under fixed AI quality. Theorem 3.1 identifies the active workflow at each skill level. We now ask where the resulting skill dynamics converge and which initial skills reach each limiting level. Let $\theta^\infty(\theta_0)$ denote the resulting long-run skill with initial skill θ_0.

Let $S_V := \{z \in R_V : f_V(z) = 0,\ f'_V(z) < 0\}$ collect stable verified-delegation roots that remain active in the dynamical system. Let B_{MV} denote switching boundaries with M immediately to the left and V immediately to the right, and define $S_{MV} := \{\tau \in B_{MV} : f_V(\tau) < 0\}$. In the finite regular case, $f_M > 0$ in the interior, so every $\tau \in S_{MV}$ satisfies $f_M(\tau) > 0 > f_V(\tau)$. Set $S := S_V \cup S_{MV}$. For analysis at switching boundaries, we use the vanishing-step limit of the conditional-mean task dynamics (Appendix C.2). The following informal statement describes the finite, nondegenerate case; see Appendices C.2 and C.5 for its proof and the general characterization, respectively.

Theorem 3.2 (**Worker convergence; informal).** *For every initial $\theta_0 \in R_P \cup R_M \cup R_V$,*

$$\theta^\infty(\theta_0) = \begin{cases} \max\{z \in S \cup \{0\} : z \le \theta_0\}, & f^\star(\theta_0) < 0, \\ \theta_0, & f^\star(\theta_0) = 0, \\ \min\{z \in S \cup \{1\} : z \ge \theta_0\}, & f^\star(\theta_0) > 0. \end{cases}$$

If $\theta_0 \in S_{MV}$, then $\theta^\infty(\theta_0) = \theta_0$ under the vanishing-step convention.

If $f^\star(\theta_0) < 0$, skill converges to the nearest attractor below θ_0, or to 0 if none exists; if $f^\star(\theta_0) > 0$, it converges to the nearest attractor above, or to 1. Adding verification therefore permits two types of interior attractor absent from the M/P benchmark: active verified-delegation equilibria in S_V and attracting manual–verified boundaries in S_{MV} (Appendix D.3).

These long-run regimes are not determined by the learning drifts alone. Proposition D.1 holds f_M, f_V, f_P fixed and varies only cost sensitivity γ. For the same initial skill $\theta_0 = 0.99$, different values of γ yield an active verified equilibrium, boundary-supported competence, or global collapse (Figure 1c). Each regime persists under sufficiently small changes in γ. Thus, sustaining skill requires not only suitable learning dynamics, but also incentives that activate them.

A complementary comparison fixes all primitives and changes only the workflow menu. Under Proposition D.2, every $\theta_0 \in (0.62, 1)$ yields limiting skill 1 under $\{M, P\}$, $\theta_V^\dagger \simeq 0.524$ under fixed V, and 0 under $\{M, V, P\}$. Adding V strictly improves optimized current utility on this interval but lowers per-task utility at limiting skill (Figure 4; Appendix D.4). Fixed verification sustains positive skill, but endogenous workflow choice abandons it before reaching its stable root.

Verification need not be dynamically self-sustaining. We can now answer the central question negatively: verification can be optimal without sustaining the skill needed to continue it.

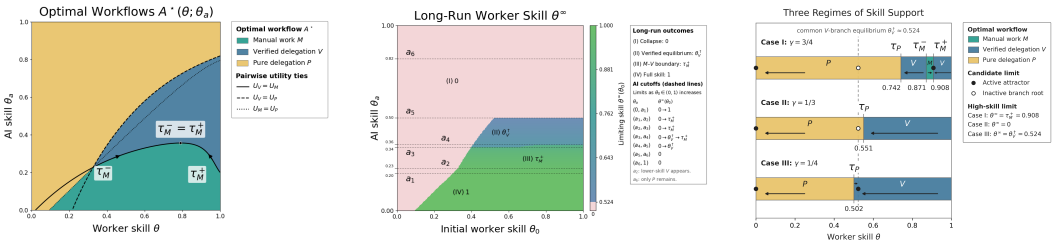
Corollary 3.3 (**Global skill collapse).** *If $R_M = \varnothing$ and $f_V(\theta) < 0$ for every $\theta \in R_V$, then $\theta^\infty(\theta_0) = 0$ for every initial $\theta_0 \in R_P \cup R_M \cup R_V$.*

The conditions remove both interior sustaining mechanisms: $R_M = \varnothing$ removes manual practice, while $f_V < 0$ on R_V rules out an active verified equilibrium. When R_V contains a high-skill region, a highly skilled worker can choose verified delegation, continue inspecting AI outputs and repairing detected errors, yet lose skill until crossing τ_P. Pure delegation then becomes optimal, and continued non-use drives skill toward zero. Skill collapse can therefore begin while human verification remains actively selected. This mechanism is absent when verification ability is fixed (Huang et al., 2026) and cannot be isolated when delegation is treated as a single action (Huang & Vishnoi, 2026). Appendix C.3 gives the proof.

Two mechanisms for sustaining positive competence. Theorem 3.2 also identifies two mechanisms that can sustain positive skill. First, a root $z \in S_V$ is an active verified-delegation equilibrium at which error-triggered repair and verification practice offset non-use decay. It prevents collapse but limits skill to $z < 1$, a *verified-delegation stagnation trap* relative to full skill under sustained manual work. Workers approaching z from above lose skill despite continued verification and repair, whereas those approaching from below gain skill.

Second, a boundary $\tau \in S_{MV}$ is supported by manual work pushing skill upward immediately below τ and verified delegation pushing it downward immediately above. Since $f_M(\tau) \neq 0$ and $f_V(\tau) \neq 0$, the limiting state is created by endogenous switching rather than by an equilibrium of either branch. Its basin can extend into the verified-delegation region above τ, so a narrow manual-work region can support positive skill over a much larger range of initial skills.

The fixed-drift comparison shows that sustaining competence is a matter of workflow incentives, not learning capacity alone. For workers seeking to preserve expertise, this suggests weighing skill



a: Optimal workflow regions. b: Limiting worker skill. c: Limits as γ varies.

Figure 1: **Workflow choice, limiting worker skill, and three regimes of skill support.** Panel a reports the current-utility-maximizing workflow as a function of worker and AI skill. Panel b reports the corresponding limiting worker skill. Panel c holds learning dynamics fixed and varies the workflow incentive γ, measuring cost sensitivity: from high interior initial skill $\theta_0 = 0.99$, the long-run outcome shifts among three possibilities; see Appendices D.1 and D.2 for details.

maintenance alongside immediate performance gains and effort savings. Independent production is therefore not merely an alternative to AI delegation: it can maintain the expertise on which effective human oversight depends.

Effects of AI quality. We now restore the dependence on θ_a to examine how AI improvement affects long-run skill. Better AI lowers both error frequency and conditional severity, increasing the attractiveness of AI workflows P and V. Meanwhile, lower error frequency reduces $q(\theta;\theta_a) = \varepsilon_a(\theta_a) r_v(\theta)$ and therefore the repair opportunities that generate a learning signal under verified delegation. The theorem below uses the finite, regular setting specified in Appendix C.4.

Theorem 3.4 (AI-quality monotonicity; informal). *For any $\theta_a' > \theta_a$ and $\theta \in (0,1)$,*

$$R_P(\theta_a') \supseteq R_P(\theta_a), \qquad R_M(\theta_a') \subseteq R_M(\theta_a), \qquad f_V(\theta;\theta_a') \le f_V(\theta;\theta_a).$$

Moreover, every regular stable verified-delegation branch $z_V(\theta_a) \in S_V(\theta_a)$ that remains active satisfies $z_V'(\theta_a) \le 0$. For every common initial skill $\theta_0 \in R_P \cup R_M \cup R_V$ at both AI qualities, $\theta^\infty(\theta_0;\theta_a') \le \theta^\infty(\theta_0;\theta_a)$.

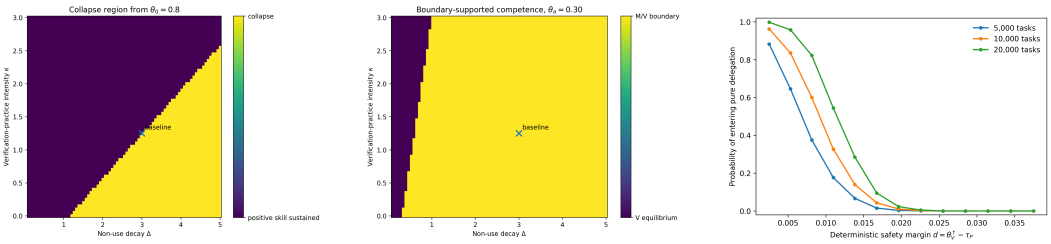
The theorem identifies two channels through which better AI can weaken human skill. On the extensive margin, pure delegation P expands and manual work M contracts; on the intensive margin, fewer AI errors reduce repair practice even when verification continues. Lower error frequency ε_a drives both channels, whereas lower conditional severity $\bar{\ell}_a$ changes workflow incentives without directly affecting repair frequency. An active stable verified equilibrium is therefore nonincreasing in AI quality, but the result extends beyond changes in a single equilibrium. Away from workflow ties, AI improvement permits only the switches $M \to V$, $M \to P$, and $V \to P$. Since $f_M \ge f_V \ge f_P$ and better AI also lowers f_V, these changes reduce the selected drift: $f^\star(\theta;\theta_a') \le f^\star(\theta;\theta_a)$. Appendix C.4 shows this drift ordering carries over to the entire skill path: from the same admissible initial skill θ_0, $\theta(t;\theta_0,\theta_a') \le \theta(t;\theta_0,\theta_a)$ for every $t \ge 0$. Under the baseline learning mechanism, better AI cannot yield higher human skill at any time, not merely at an unchanged equilibrium. Taking $t \to \infty$ gives the long-run comparison even when the identity of the attractor changes.

Parametric illustration. Figure 1 compares workflow choice and long-run skill across worker skill, AI quality, and cost sensitivity. Appendix D gives the specifications and derivations.

AI use can be compensatory and complementary (Figure 1a). At $\theta_a = 0.30$, the workflow ordering is $P \to V \to M \to V$. Verified delegation therefore can both compensate for weak production at lower skill and complement reliable detection and repair at higher skill.

A small manual region can have a large dynamic effect (Figures 1a and 1b). At $\theta_a = 0.356$, manual work is optimal only on approximately $(0.775, 0.803)$, an interval of width 0.028. Yet every worker with $\theta_0 \gtrsim 0.775$, about 22% of the initial-skill range, converges to the boundary-supported level 0.803. The manual region stops further skill loss among workers approaching from the higher-skill verified region, so its effect extends beyond the skills at which manual work is initially optimal.

Sustainable skill disappears before verification does (Figure 1b). As θ_a rises from 0.40 to 0.50, the verified equilibrium falls from 0.604 to 0.524, while its basin contracts from $(0.439, 1)$ to $(0.521, 1)$. Above $\theta_a \simeq 0.502$, τ_P passes the still-positive verified equilibrium, making it inactive and causing



a: Sustainable V vs. collapse b: Boundary-supported skill c: Finite-task escape

Figure 2: **Robustness.** (a) At $\theta_a = 0.5$, varying (κ, Δ) yields sustainable verified delegation or collapse. (b) At $\theta_a = 0.3$, boundary-supported competence persists over a broad parameter region. (c) Starting at the stable V equilibrium, escape into pure delegation increases as the safety margin shrinks and exposure grows. Crosses mark $(\kappa, \Delta) = (1.25, 3)$.

global collapse. At $\theta_a = 0.51$, even workers starting at $\theta_0 = 0.99$ initially choose V, lose skill until entering P, and then converge to zero. Yet verification remains optimal over a high-skill region until $\theta_a \simeq 0.819$. Thus, high initial skill and an initially optimal verified workflow do not guarantee sustained human capability.

Greater cost sensitivity need not lower long-run skill (Figure 1c). With all other primitives fixed, increasing γ across the illustrated cases changes the long-run outcome from a verified equilibrium near 0.524, to collapse, and then to high skill near 0.908 sustained by manual practice. Greater cost sensitivity can therefore favor skill-preserving manual work, rather than simply deepen AI reliance. For institutions seeking to maintain expertise, this suggests evaluating efficiency incentives by the opportunities for independent practice they preserve, rather than assuming that stronger cost pressure necessarily accelerates deskilling (Appendix D.2).

4 Robustness and Empirical Interpretation

The qualitative conclusions of the model raise three questions about robustness and empirical interpretation. Do collapse, verified-delegation equilibria, and boundary-supported competence persist beyond the baseline learning parameters? Does a stable mean-field equilibrium remain durable when learning opportunities occur randomly at the task level? And how can observations from AI-assisted workflows be mapped to model inputs? We address these questions in turn; Appendix E provides the robustness analyses, and Appendix F illustrates the data-to-parameter mapping.

Robustness to human learning parameters. The parametric illustration fixes $\kappa = 5/4$ and $\Delta = 3$. We vary $\kappa \in [0, 3]$ and $\Delta \in [0.1, 5]$ while holding all current-utility primitives fixed. Because neither parameter enters current workflow utility, the workflow partition remains unchanged within each exercise, isolating the learning side of the model.

At $\theta_a = 0.5$, the worker chooses pure delegation P at low skill and verified delegation V at higher skill. Figure 2a shows substantial regions of collapse and sustainable verified delegation from $\theta_0 = 0.8$. Stronger verification practice or weaker non-use decay supports an active interior verified-delegation equilibrium; sufficiently strong decay relative to verification practice moves that equilibrium below the P/V switching boundary, after which the trajectory enters pure delegation P and converges toward zero. At the baseline, $\tau_P \simeq 0.521$, $\theta_V^\dagger \simeq 0.524$, so verified delegation V remains deterministically sustainable, but with only a small positive safety margin.

The second panel considers $\theta_a = 0.3$, where the workflow ordering is $P \to V \to M \to V$. Starting from $\theta_0 = 0.95$, the baseline dynamics converge to the upper manual–verified boundary, $\tau_M^+ \simeq 0.935$. Figure 2b shows that this boundary-supported regime persists over a broad region of (κ, Δ); elsewhere, the upper verified branch supports an interior equilibrium. In the boundary-supported region, manual practice prevents the further skill loss that verified delegation would induce. Thus collapse, verified-delegation stagnation, and boundary-supported competence each persist over nontrivial regions of the learning parameters; see Appendix E.1 for details.

Beyond the verification-practice signal already present on every reviewed task, we allow studying correct AI outputs to provide an additional learning signal of strength $\xi(1 - \theta)$, $\xi \geq 0$. At fixed AI quality, this channel raises the verified-delegation drift without changing current workflow in-

centives. Over the same (κ, Δ) grid, increasing ξ from 0 to 1 reduces the fraction of parameter combinations leading to collapse from 33.9% to 22.2%. Thus, the collapse mechanism can persist even with additional learning from correct AI outputs; Appendix E.2 gives the analysis.

Finite-task dynamics. The deterministic model averages over the discrete occurrence of AI errors and successful detections. Under verified delegation, generative repair occurs with probability $q(\theta; \theta_a) = \varepsilon_a(\theta_a) r_v(\theta)$. We therefore construct a task-level stochastic realization in which errors and detections occur randomly and workflow choice is recomputed after every task. Its conditional expected skill increment equals $\eta f_A(\theta; \theta_a)$ whenever projection is inactive. Random skill updates can then change later workflow choices under the same utility-maximizing rule. Let $d = \theta_V^\dagger - \tau_P$ denote the safety margin of a stable verified-delegation equilibrium. We initialize workers at $\theta_V^\dagger \simeq 0.524$ and vary γ, which moves τ_P without changing the verified-delegation learning branch. For each margin, we simulate $5{,}000$ independent task histories with per-task step size $\eta = 5 \times 10^{-4}$.

Figure 2c shows that escape into pure delegation becomes more likely both as the safety margin shrinks and as exposure grows. With $d \simeq 1.39 \times 10^{-2}$, escape probabilities are 6.7%, 14.0%, and 28.5% after 5,000, 10,000, and 20,000 tasks; with $d \simeq 2.5 \times 10^{-3}$, they rise to 88.2%, 96.3%, and 99.8%. The mean drift is locally restoring near the verified equilibrium, but a downward crossing of τ_P makes pure delegation optimal, after which its negative drift pushes skill farther away. Seed and step-size checks give the same qualitative pattern. Thus, *mean stability need not imply finite-task durability*.

Illustrative data-to-parameter mapping. Appendix F illustrates how performance and interaction records from Bastani et al. (2025) map to selected model inputs under explicit workflow proxies and normalizations. The exercise also identifies quantities that require additional measurements. It is not an empirical validation of the proposed mechanism or its long-run predictions.

5 Conclusion, Limitations, and Implications

We study whether human verification remains dynamically sustainable when the competence supporting it is itself shaped by repeated workflow choice. The feedback between current workflow choice and future skill produces collapse, interior verified-delegation equilibria, and competence sustained at manual–verified switching boundaries. Most importantly, verified delegation can remain optimal over a nonempty skill range even when no positive skill level is sustainable. Verification can continue without enough production and repair practice to maintain skill, while even a narrow region of manual work can replenish skill later used under verified delegation. Better AI can further reduce opportunities for detection and repair. Whether skill is sustained depends on learning and decay rates, reviewer reliability, task structure, and workflow costs.

The model is intentionally stylized. Human competence is low-dimensional, AI capability is fixed over the learning horizon, workflow choice is myopic with respect to future skill, and the data-to-parameter mapping is illustrative rather than a joint structural estimation: treatments serve as workflow proxies, and several inputs require additional assumptions or measurements. The robustness exercises show that the deterministic regimes persist across broad learning-parameter ranges and with learning from correct reviewed outputs. Finite-task simulations further show that random repair opportunities can undermine a mean-stable verified equilibrium.

A natural next step is to let production and verification skills evolve separately and ask whether verification remains effective as production skill declines. Other extensions could allow AI capability and task allocation to evolve endogenously, verification difficulty and learning opportunities to vary with error type or severity, or workers and institutions to internalize the future value of human competence. Relaxing calibrated beliefs would also allow over- or under-reliance to affect the same feedback. Empirically, the key next step is to measure workflow choice, error detection and repair, and subsequent unaided skill over repeated episodes. Such data would allow separate estimation of learning and decay parameters and tests of the predicted workflow boundaries and long-run regimes.

Our results suggest evaluating AI-assisted workflows not only by current performance and effort savings, but also by the learning opportunities they preserve. For workers and institutions seeking durable oversight, the goal is not merely to retain a human reviewer, but to sustain the expertise that makes review effective.

AI USE STATEMENT

The authors developed the model, main results, and proof arguments. Generative AI tools assisted with language editing, proof checking and the drafting of selected proof details, as well as coding and plotting. The authors reviewed the AI-assisted material and take responsibility for the final content of the paper.

ETHICS STATEMENT

This work consists of theoretical analysis, simulations, and secondary analysis of publicly released data. We conduct no new human-subject experiments and collect no new personal data.

REPRODUCIBILITY STATEMENT

Model derivations are provided in Appendix B, and complete proofs of the main theoretical results are provided in Appendix C. Appendices D–E specify the numerical parameterizations, robustness analyses, and simulation procedures. Appendix F describes the data sources, workflow proxies, normalizations, and parameter mapping used in the illustrative empirical analysis. Code for reproducing the numerical results, figures, and data analyses is provided in the supplementary material.

REFERENCES

Daron Acemoglu, Dingwen Kong, and Asuman Ozdaglar. AI, human cognition and knowledge collapse. Working Paper 34910, National Bureau of Economic Research, February 2026.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor (eds.), *Séminaire de Probabilités XXXIII*, pp. 1–68, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48407-3.

Robert Bergman. The meaning—and illusion—of human oversight of AI. Mediate.com, 2025. URL https://mediate.com/the-meaning-and-illusion-of-human-oversight-of-ai/.

Robert A. Bjork. Memory and metamemory considerations in the training of human beings. In Janet Metcalfe and Arthur P. Shimamura (eds.), *Metacognition: Knowing About Knowing*, pp. 185–205. MIT Press, 1994.

Jamie I. D. Campbell. Production, verification, and priming of multiplication facts. *Memory & Cognition*, 15(4):349–364, 1987. doi: 10.3758/BF03197037.

Jamie I. D. Campbell and Derek P. M. Tarling. Retrieval processes in arithmetic production and verification. *Memory & Cognition*, 24(2):156–172, 1996. doi: 10.3758/BF03200878.

Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132 (3):354–380, 2006. doi: 10.1037/0033-2909.132.3.354.

Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182, 1989. doi: 10.1207/s15516709cog1302_1. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1302_1.

Fabrizio Dell'Acqua, Edward McFowland, III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, 2023. URL https://www.hbs.edu/ris/download.aspx?name=24-013.pdf.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000. doi: 10.4324/9781410605269.

Krzysztof Z. Gajos and Lena Mamykina. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *IUI*, pp. 794–806. ACM, 2022.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):31, 2021.

Sandra Grinschgl, Frank Papenmeier, and Hauke S Meyerhoff. Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, 74(9):1477–1496, 2021.

Cornelia S. Große and Alexander Renkl. Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, 17(6):612–634, 2007. doi: 10.1016/j.learninstruc.2007.09.008.

Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 221–236. Association for Computing Machinery, 2024. doi: 10.1145/3630106.3658901. URL https://doi.org/10.1145/3630106.3658901.

Andrew Heathcote, Scott Brown, and D. J. K. Mewhort. The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7(2):185–207, Jun 2000. ISSN 1531-5320. doi: 10.3758/BF03212979. URL https://doi.org/10.3758/BF03212979.

Lingxiao Huang and Nisheeth K. Vishnoi. Path dependence under adaptive AI delegation. *arXiv preprint arXiv:2603.02950*, 2026. doi: 10.48550/arXiv.2603.02950. URL https://arxiv.org/abs/2603.02950.

Lingxiao Huang, Wenyang Xiao, and Nisheeth K. Vishnoi. Delegation and verification under AI. *arXiv preprint arXiv:2603.02961*, 2026. doi: 10.48550/arXiv.2603.02961. URL https://arxiv.org/abs/2603.02961.

Jeffrey D. Karpicke and Henry L. Roediger. The critical importance of retrieval for learning. *Science*, 319(5865):966–968, 2008. doi: 10.1126/science.1152408.

Bernard O Koopman. The theory of search. II. Target detection. *Operations Research*, 4(5):503–531, 1956.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv preprint arXiv:2506.08872, 2025. URL https://arxiv.org/abs/2506.08872.

Grace Liu, Brian Christian, Tsvetomira Dumbalska, Michiel A. Bakker, and Rachit Dubey. AI assistance reduces persistence and hurts independent performance. *arXiv preprint arXiv:2604.04721*, 2026. doi: 10.48550/arXiv.2604.04721. URL https://arxiv.org/abs/2604.04721.

David Lyell and Enrico Coiera. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, 2017.

Matthew P. McCurdy, Wolfgang Viechtbauer, Allison M. Sklenar, Andrea N. Frankenstein, and Eric D. Leshikar. Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, 27(6):1139–1165, 2020. doi: 10.3758/s13423-020-01762-3.

Bruce M. McLaren, Deanne Adams, Kelley Durkin, George Goguadze, Richard E. Mayer, Bethany Rittle-Johnson, Sergey Sosnovsky, Seiji Isotani, and Martin van Velsen. To err is human, to explain and correct is divine: A study of interactive erroneous examples with middle school math students. In *21st Century Learning for 21st Century Skills*, volume 7563 of *Lecture Notes in Computer Science*, pp. 222–235. Springer, 2012. doi: 10.1007/978-3-642-33263-0_18. URL https://www.cs.cmu.edu/~bmclaren/pubs/McLarenEtAl-AdaptErrExStudy-ECTEL2012.pdf.

Allen Newell and Paul S. Rosenbloom. Mechanisms of skill acquisition and the law of practice. In John R. Anderson (ed.), *Cognitive Skills and Their Acquisition*, pp. 1–55. Lawrence Erlbaum, Hillsdale, NJ, 1981.

Gali Noti, Kate Donahue, Jon M. Kleinberg, and Sigal Oren. AI-assisted decision making with human learning. In *EC*, pp. 38–39. ACM, 2025.

R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000. doi: 10.1109/3468.844354.

Samir Passi and Mihaela Vorvoreanu. Overreliance on AI literature review. Technical Report MSR-TR-2022-12, Microsoft Research, June 2022. URL https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/.

Philip I. Pavlik, Jr. and John R. Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29(4):559–586, 2005. doi: 10.1207/s15516709cog0000_14. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_14.

Kenny Peng, Nikhil Garg, and Jon M. Kleinberg. A no free lunch theorem for human-AI collaboration. In *AAAI*, pp. 14369–14376. AAAI Press, 2025.

Adam A. Porter and Lawrence G. Votta. Comparing detection methods for software requirements inspections: A replication using professional subjects. *Empirical Software Engineering*, 3(4):355–379, 1998. doi: 10.1023/A:1009776104355.

R. A. Rescorla and A. R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy (eds.), *Classical Conditioning II: Current Research and Theory*, pp. 64–99. Appleton-Century-Crofts, New York, 1972.

Evan F Risko and Sam J Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016. doi: 10.1016/j.tics.2016.07.002.

Henry L. Roediger and Jeffrey D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006. doi: 10.1111/j.1467-9280.2006.01693.x.

Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*, 2026. doi: 10.48550/arXiv.2601.20245. URL https://arxiv.org/abs/2601.20245.

Norman J. Slamecka and Peter Graf. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6):592–604, 1978. doi: 10.1037/0278-7393.4.6.592.

Nenad Tomašev, Matija Franklin, and Simon Osindero. Intelligent AI delegation. *arXiv preprint arXiv:2602.11865*, 2026. doi: 10.48550/arXiv.2602.11865. URL https://arxiv.org/abs/2602.11865.

N. Jane Zbrodoff and Gordon D. Logan. On the relation between production and verification tasks in the psychology of simple arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1):83–97, 1990. doi: 10.1037/0278-7393.16.1.83.

CONTENTS

1 Introduction **1**

2 Model **3**

3 Theoretical results **5**

4 Robustness and empirical interpretation **8**

5 Conclusion, limitations, and implications **9**

A Additional related work **14**

B Details from Section 2: model and skill dynamics **15**

B.1 Notation and workflows 16

B.2 Expected workflow losses and costs 16

B.3 From task-level learning to skill dynamics 19

B.4 Connection to the static delegation and verification model 21

B.5 Connection to the adaptive AI-reliance model 22

C Proofs and additional results from Section 3 **22**

C.1 Proof of Theorem 3.1: optimal workflow choice 23

C.2 Proof of Theorem 3.2: skill convergence and attraction basins 24

C.3 Proof of Corollary 3.3: conditions for global skill collapse 26

C.4 Proof of Theorem 3.4: effects of AI quality 27

C.5 Degenerate cases and general vanishing-step limits 28

D Parametric illustrations and counterfactual comparisons **29**

D.1 Baseline specification and parametric analysis 29

D.1.1 Primitives and workflow regions 29

D.1.2 Skill dynamics and critical AI qualities 30

D.2 Details from Figure 1c: limit shifts by workflow incentives 32

D.3 Comparing dynamics with and without verification 34

D.4 Short-term utility gains, long-term skill loss 35

E Details from Section 4: learning robustness and finite-task dynamics **37**

E.1 Effects of learning and decay rates 38

E.2 Additional learning from reviewing correct AI outputs 39

E.3 Stochastic escape into pure delegation 40

F Details from Section 4: illustrative data-to-parameter mapping **42**

Guide to the appendices. Appendix A locates the contribution in the related literature. Appendix B defines the task-level objects and distinguishes realized updates, conditional-mean updates, and continuous-time limits. Appendix C states the precise conditions and proofs corresponding to the main-text theorems and corollary; its last subsection treats degenerate cases separately. Appendix D first specifies the common numerical baseline and then uses two controlled comparisons to isolate workflow incentives and the availability of verification. Appendix E studies learning-parameter robustness and finite-task escape under that baseline. Finally, Appendix F separates observable counterparts, reported conditional fits, and imposed inputs in an illustrative data mapping. The numerical examples and empirical mapping are not prerequisites for the general proofs.

A Additional related work

Our work connects research on human–AI delegation and verification, dynamic effects of AI reliance on human skill, empirical studies of AI-assisted learning, and cognitive mechanisms underlying generation and review. The distinctive question is whether a workflow that retains human verification also preserves the competence required to perform that verification over time.

Human–AI delegation and verification. A growing literature studies how tasks should be divided between humans and AI systems, including work on delegation, complementary expertise, and joint decision-making (Tomašev et al., 2026; Peng et al., 2025). Most closely related is the composer–reviewer framework of Huang et al. (2026), which separates execution from verification and characterizes when a worker chooses manual work, verified delegation, or pure delegation as a function of fixed execution and verification abilities. That framework shows that verification reliability can make AI-assisted work preferable to both independent production and unchecked delegation, and that small differences in worker abilities can induce sharp changes in workflow choice.

The present paper asks whether the abilities supporting those workflows are maintained through repeated use. The static framework characterizes which workflow is optimal for fixed worker abilities; here, workflow choice changes latent skill and hence future production and verification reliabilities. Thus the object of interest is not only whether verification is valuable at a given level of competence, but whether repeated verified delegation maintains the competence on which its value depends.

AI reliance and skill dynamics. A complementary literature studies how repeated reliance on AI or automation can alter human skill and incentives to acquire or maintain human capital. The dynamic framework of Huang & Vishnoi (2026) models feedback between AI reliance and human skill, showing that reduced practice can make delegation more attractive and generate multiple equilibria, path dependence, and long-run skill degradation. Related economic work studies how agentic AI can weaken learning incentives and erode the shared stock of general knowledge (Acemoglu et al., 2026).

Our model decomposes delegated work into generation, verification, and repair. In the baseline, verified delegation is not equivalent to partial reliance across tasks: a worker who manually performs a fraction of tasks receives generative practice on each of those tasks. By contrast, a worker who reviews every AI-generated output receives generative practice only when an AI error occurs, is detected, and is repaired. These within-task learning channels, coupled with utility-based workflow choice, yield two verification-specific outcomes: active stable verified-delegation equilibria and competence sustained at attracting manual–verified switching boundaries (Theorem 3.2). These mechanisms arise from a separate verification workflow with error-triggered repair and switching to manual work, which Huang & Vishnoi (2026) does not model. The model also makes possible the central phenomenon studied here: human skill can erode while verification remains actively selected.

Taken together, these two frameworks separate the two ingredients combined here: Huang et al. (2026) distinguishes generation from verification while holding human abilities fixed, whereas Huang & Vishnoi (2026) endogenizes human skill while treating delegation as an atomic action. The present model combines these dimensions by making the competence underlying verified delegation endogenous.

Assisted performance and independent skill. A growing empirical literature documents a distinction between performance with AI assistance and subsequent independent capability. AI assistance can improve immediate task performance, speed, or completion while producing weaker retention, transfer, conceptual understanding, or later unaided performance (Liu et al., 2026; Shen & Tamkin, 2026; Kosmyna et al., 2025; Bastani et al., 2025). Learning outcomes also vary with how assistance is used: cognitive engagement, explanation-seeking, and active problem solving are associated with stronger learning than reliance on completed solutions or suggestions (Liu et al., 2026; Shen & Tamkin, 2026; Gajos & Mamykina, 2022).

Our framework models one mechanism through which assisted performance and latent capability can diverge. Verified delegation can maintain strong current performance because a sufficiently skilled worker continues to detect and repair AI errors, even while the workflow supplies too little practice to maintain that skill in the long run. The model therefore distinguishes the current value of human oversight from the dynamic sustainability of the capability supporting it.

Learning through generation, verification, and repair. The learning mechanism in our model draws on several established findings in cognitive science. The generation effect documents that actively producing information can yield stronger learning than receiving it (Slamecka & Graf, 1978; McCurdy et al., 2020). Research comparing arithmetic production and verification tasks suggests that the two are distinct but draw on overlapping knowledge and cognitive processes (Campbell, 1987; Zbrodoff & Logan, 1990; Campbell & Tarling, 1996). Related work shows that retrieval and continued practice strengthen retention (Roediger & Karpicke, 2006; Karpicke & Roediger, 2008; Cepeda et al., 2006), while detecting and correcting errors in worked examples can itself generate learning (Große & Renkl, 2007; McLaren et al., 2012).

At the same time, work on cognitive offloading shows that external assistance can improve current performance while reducing the internal cognitive activity or memory involved in completing the task (Risko & Gilbert, 2016; Grinschgl et al., 2021). These findings motivate, rather than constitute, our learning law. We do not assume that verification produces no learning. Instead, we allow inspection to generate a positive practice signal while distinguishing it from the additional generative practice associated with production and repair. Our contribution is not to establish this cognitive asymmetry, but to characterize its consequences when workflow choice and human competence co-evolve.

Learning-aware AI assistance. Some theoretical work asks how assistance can be designed explicitly to promote future human learning rather than only current performance. For example, Noti et al. (2025) study a forward-looking algorithm that selects what information to reveal to a human learner and may sacrifice short-run assistance in order to improve long-run understanding. This perspective is complementary to ours. In our baseline, the worker chooses a workflow to maximize current utility rather than to preserve future competence. The resulting distinction is important: AI assistance can in principle be designed to sustain human learning, but merely retaining a human reviewer does not guarantee that ordinary workflow choice will do so.

Role of the supplementary analyses. These connections motivate the distinction between production, verification, and repair; they do not identify the numerical learning law or establish its long-run predictions empirically. Appendix B next makes the probabilistic construction and the pointwise correspondences to the prior models explicit. The formal dynamic conclusions are proved separately in Appendix C.

B Details from Section 2: model and skill dynamics

This appendix derives expected workflow losses and costs from task-level events and constructs bounded skill updates with the conditional mean drift in (5). The last two subsections relate these payoffs and learning drifts to the preceding models. Appendix C.2 specifies the dynamics at workflow ties.

B.1 NOTATION AND WORKFLOWS

At each task, the worker uses calibrated expectations of final loss and completion cost to choose $A^* \in \arg\max_{A \in \mathcal{A}} U_A$, with a fixed tie rule. The chosen workflow determines task outcomes and learning exposure; future learning does not enter the current objective. Figure 3 illustrates this sequence. Uppercase M, V, P denote workflows; subscripts m, v distinguish manual-output and verification reliability, while g, s, v distinguish generation, self-checking, and AI-output inspection costs.

Table 1 collects notation used in the main text; task-level variables needed for the derivations are introduced below. The strict regions R_A exclude utility ties: below τ_P, pure delegation is weakly optimal, not necessarily uniquely optimal. The drift f_A describes the conditional mean rate of skill change under a fixed workflow, not a realized task outcome; Appendix B.3 derives this distinction.

Table 1: Selected notation used in the main text.

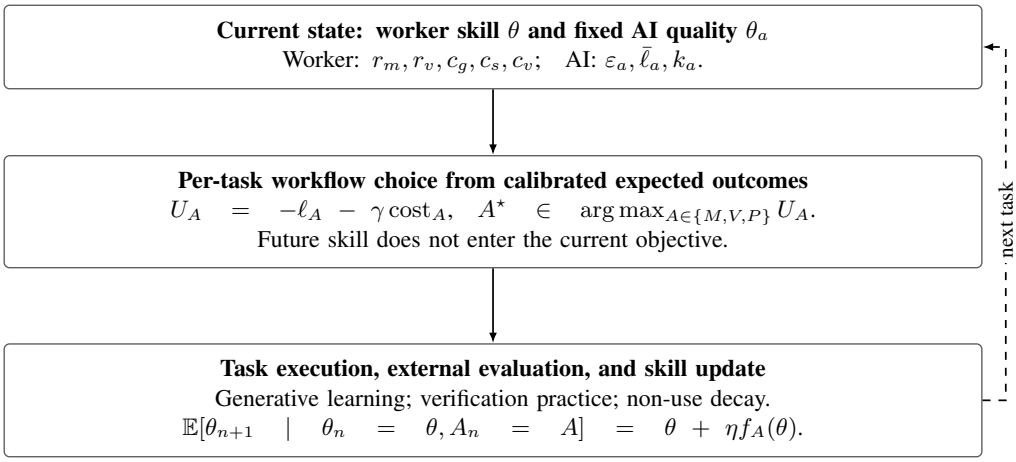
Symbol	Meaning
$t;\ \theta(t), \theta_0$	Task-normalized time; current and initial worker skill in $[0, 1]$.
$\theta_a \in [0, 1]$	Fixed task-specific AI-quality index.
$\mathcal{A} = \{M, V, P\}$	Manual work, verified delegation, pure delegation.
$A;\ A^*(\theta; \theta_a)$	A workflow; a current-utility maximizer selected by the fixed tie rule.
$r_m(\theta), r_v(\theta)$	Manual-output reliability; conditional error-detection probability.
ε_a, ℓ_a	AI-error probability and conditional mean severity.
$q = \varepsilon_a r_v$	Probability of an AI error being detected and repaired.
ℓ_A	Expected final loss under A; $\ell_M = 1 - r_m$.
c_g, c_s, c_v	Generation, self-checking, and AI-output inspection costs.
k_j, b_j, d	Floor, scale, and exponent in $c_j = k_j + b_j(1-\theta)^d$.
$k_a;\ \lambda \in [0, 1]$	AI-use cost; repair effort relative to generation.
cost_A	Expected completion cost under A.
$\gamma \geq 0;\ U_A$	Cost weight; per-task utility $-\ell_A - \gamma\,\mathrm{cost}_A$.
$\rho;\ p_m, p_v$	Task tractability; exponents in the illustrative reliability family.
$\Delta > 0, \kappa \geq 0$	Non-use decay and verification-practice rates.
$f_A;\ f^\star$	Fixed-workflow skill drift; active drift selected by current utility.
$G;\ D$	Utility gaps $\max\{U_M, U_V\} - U_P$ and $U_V - U_M$.
$R_A(\theta_a)$	Skills at which A is strictly optimal; utility ties are excluded.
$\tau_P;\ \tau^{(1)}, \ldots, \tau^{(n)}$	Cutoff above which P is strictly dominated; sign-changing M/V crossings above that cutoff.
$\sigma \in \{-1, +1\}$	Sign of D immediately above τ_P in the finite-crossing case.
$S_V;\ S_{MV}$	Active verified roots with $f_V' < 0$; attracting $M \to V$ switching boundaries.
$z_V(\theta_a)$	A regular branch of stable verified roots that remain active as AI quality changes.
$\theta_V^\dagger$	A fixed point of the verified-delegation branch, when defined; not necessarily active.
$\theta^\infty(\theta_0; \theta_a)$	Long-run limit shared by all vanishing-step subsequential trajectories, only when their limits exist and agree (Appendices C.2 and C.5).

B.2 EXPECTED WORKFLOW LOSSES AND COSTS

Fix (θ, θ_a) and condition on this pre-task state throughout the subsection; we suppress that conditioning in the notation. All loss variables are integrable. Conditional moments given an event are required only where that event has positive probability. An event-weighted expectation is zero when the event has probability zero, so the formulas also cover perfect AI and zero detection probability.

Performance loss is assigned to the submitted output by an external evaluator. The evaluator is not a strategic player; evaluation and feedback reveal task outcomes, while the skill response to that feedback is specified separately in Appendix B.3. Completion costs measure the worker's activities, not the evaluator's loss.

Potential task events. For comparing workflows before one is chosen, define $E \in \{0, 1\}$ as the error indicator of a potential AI draft, and $I_D \in \{0, 1\}$ as the detection outcome that would occur if that draft were reviewed. These potential outcomes are defined whether or not AI or review is



(a) Current incentives select a workflow; its learning channels update skill.

M: produce manually and self-check

$$\ell_M = 1 - r_m, \qquad \mathrm{cost}_M = c_g + c_s$$

Learning: generation and checking; no non-use decay.

V: verify the AI draft, then repair detected errors

Three mutually exclusive outcomes; $q = \varepsilon_a r_v$.

Outcome	Probability	Mean loss	Total cost
Detected error	q	ℓ_M	$k_a + c_v + \lambda c_g$
Missed error	$\varepsilon_a(1 - r_v)$	$\bar{\ell}_a$	$k_a + c_v$
Correct draft	$1 - \varepsilon_a$	0	$k_a + c_v$

$$\ell_V = (\varepsilon_a - q)\bar{\ell}_a + q\ell_M, \qquad \mathrm{cost}_V = k_a + c_v + \lambda q c_g$$

Learning: repair with probability q; non-use otherwise; checking in both.

P: accept the AI draft without checking

$$\ell_P = \varepsilon_a \bar{\ell}_a, \qquad \mathrm{cost}_P = k_a$$

Learning: non-use decay only, whether the AI draft is correct or not.

(b) Workflow outcomes and the corresponding learning channels.

Figure 3: **Workflow under AI assistance.** Skill and AI-quality arguments are suppressed. Losses in panel (b) are conditional means, not realized task losses; panel (a) shows the conditional expected update for sufficiently small η. Appendices B.2 and B.3 derive the outcome averages and learning dynamics.

ultimately selected. Set $I_D = 0$ when $E = 0$ to rule out false positives; do not set the potential I_D to zero merely because the chosen workflow is P. The probability primitives are

$$\mathbb{P}(E = 1) = \varepsilon_a, \qquad \mathbb{P}(I_D = 1 \mid E = 1) = r_v(\theta).$$

Consequently,

$$\begin{aligned}\mathbb{P}(E = 1, I_D = 1) &= \mathbb{P}(E = 1)\mathbb{P}(I_D = 1 \mid E = 1)\\ &= \varepsilon_a r_v(\theta) =: q,\\ \mathbb{P}(E = 1, I_D = 0) &= \mathbb{P}(E = 1)\mathbb{P}(I_D = 0 \mid E = 1)\\ &= \varepsilon_a(1 - r_v(\theta)).\end{aligned}$$

Thus the three paths under V are a correct AI output, an undetected error, and a detected error followed by repair, with probabilities $1 - \varepsilon_a$, $\varepsilon_a(1 - r_v(\theta))$, and q, respectively. The multiplication above is the conditional-probability rule; it does not require independence of E and I_D.

Manual work and pure delegation. Let Ξ_M be the evaluated loss of the worker's final manual output, including ordinary self-checking. The mean primitive is $\mathbb{E}[\Xi_M] = \ell_M(\theta) = 1 - r_m(\theta)$. Let Ξ_a be the evaluated severity of an AI error, with $\mathbb{E}[\Xi_a \mid E = 1] = \bar{\ell}_a$, and extend Ξ_a by zero when $E = 0$. The potential realized losses are $L_M = \Xi_M$ and $L_P = E\Xi_a$. Hence

$$\begin{aligned}\mathbb{E}[L_M] &= \mathbb{E}[\Xi_M] = \ell_M(\theta),\\ \mathbb{E}[L_P] &= \mathbb{E}[E\Xi_a]\\ &= \mathbb{P}(E = 1)\mathbb{E}[\Xi_a \mid E = 1] + 0\\ &= \varepsilon_a\bar{\ell}_a = \ell_P.\end{aligned}$$

This recovers (2) without identifying error frequency with conditional error severity.

Verified delegation. The separable baseline detection mechanism can be expressed as

$$\mathbb{P}(I_D = 0 \mid E = 1, \Xi_a) = 1 - r_v(\theta) :$$

conditional on an AI error and the fixed pre-task state, the probability of missing that error does not depend on its severity. By iterated expectations,

$$\begin{aligned}\mathbb{E}[(1 - I_D)\Xi_a \mid E = 1] &= \mathbb{E}[\mathbb{E}[(1 - I_D)\Xi_a \mid E = 1, \Xi_a] \mid E = 1]\\ &= \mathbb{E}[\Xi_a\mathbb{P}(I_D = 0 \mid E = 1, \Xi_a) \mid E = 1]\\ &= (1 - r_v(\theta))\mathbb{E}[\Xi_a \mid E = 1]\\ &= (1 - r_v(\theta))\bar{\ell}_a.\end{aligned}$$

Equivalently for the loss calculation, it suffices to impose only

$$\mathbb{E}[\Xi_a \mid E = 1, I_D = 0] = \bar{\ell}_a$$

when missed errors have positive probability. Conditional independence $I_D \perp \Xi_a \mid (E = 1, \theta, \theta_a)$ implies this mean restriction but is stronger than necessary.

Let Ξ_{rep} be post-repair loss. The manual-repair convention is

$$\mathbb{E}[\Xi_{\text{rep}} \mid E = 1, I_D = 1] = \ell_M(\theta).$$

It matches mean repair quality to mean manual-output quality; it does not assert full independence of repair outcomes and the original error. Extend Ξ_{rep} by zero outside the repair event. Then

$$L_V = E(1 - I_D)\Xi_a + EI_D\Xi_{\text{rep}}.$$

Computing the two terms separately gives

$$\begin{aligned}\mathbb{E}[E(1 - I_D)\Xi_a] &= \mathbb{P}(E = 1)\mathbb{E}[(1 - I_D)\Xi_a \mid E = 1]\\ &= \varepsilon_a(1 - r_v(\theta))\bar{\ell}_a,\\ \mathbb{E}[EI_D\Xi_{\text{rep}}] &= \mathbb{P}(E = 1, I_D = 1)\mathbb{E}[\Xi_{\text{rep}} \mid E = 1, I_D = 1]\\ &= q\ell_M(\theta).\end{aligned}$$

Linearity of expectation therefore yields

$$\begin{aligned}\mathbb{E}[L_V] &= \mathbb{E}[E(1 - I_D)\Xi_a] + \mathbb{E}[EI_D\Xi_{\text{rep}}]\\ &= \varepsilon_a(1 - r_v(\theta))\bar{\ell}_a + q\ell_M(\theta) = \ell_V,\end{aligned}$$

which is (3). Correct AI outputs contribute zero.

Completion costs. For this task-level realization, activity costs are deterministic conditional on the current state. Completion cost is nevertheless random under V because repair is incurred only after a detected error. Denote realized completion cost by C_A locally in this subsection:

$$C_M = c_g + c_s, \qquad C_P = k_a, \qquad C_V = k_a + c_v + EI_D\lambda c_g.$$

Using that k_a, c_g, c_s, c_v, and λ are fixed under the present conditioning,

$$\begin{aligned}
\mathbb{E}[C_M] &= \mathbb{E}[c_g + c_s] = c_g + c_s = \mathrm{cost}_M, \\
\mathbb{E}[C_P] &= \mathbb{E}[k_a] = k_a = \mathrm{cost}_P, \\
\mathbb{E}[C_V] &= \mathbb{E}[k_a + c_v + EI_D\lambda c_g] \\
&= k_a + c_v + \lambda c_g \mathbb{E}[EI_D] \\
&= k_a + c_v + \lambda c_g \mathbb{P}(E = 1, I_D = 1) \\
&= k_a + c_v + \lambda q c_g = \mathrm{cost}_V\,.
\end{aligned}$$

Thus (4) holds without an independence assumption between random repair effort and detection: there is no additional random activity-cost variable in this realization. Allowing such randomness would instead require specifying its conditional mean on the repair event. There is no extra c_s charge under V: c_v covers inspection, and the baseline contains no additional post-repair self-check. The factor λ discounts repair implementation effort, not repair quality or learning exposure.

From potential outcomes to current workflow choice. All three payoff pairs (L_A, C_A) have been defined before selecting a workflow; only the chosen pair is used on the actual task. For every $A \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{E}[-L_A - \gamma C_A] &= -\mathbb{E}[L_A] - \gamma \mathbb{E}[C_A] \\
&= -\ell_A - \gamma\,\mathrm{cost}_A = U_A.
\end{aligned}$$

Calibrated, pre-realization choice therefore maximizes the expected utility of each candidate workflow. This is $\max_A \mathbb{E}[-L_A - \gamma C_A]$, not the generally different quantity $\mathbb{E}[\max_A\{-L_A - \gamma C_A\}]$, which would allow a choice after observing all potential outcomes.

B.3 From task-level learning to skill dynamics

Following Huang & Vishnoi (2026), we specify a task-level learning law and average its skill increment conditional on the pre-task state. The learning signals and bounded update geometry are modeling choices. In particular, the equality $\mathbb{E}[L_M] = \ell_M(\theta)$ does not itself imply that learning follows $-\ell'_M(\theta)$. Evaluation supplies task feedback; the following law specifies how active production, detection, repair, and inspection convert that feedback into skill changes.

Learning on an individual task. Index tasks by $n = 0, 1, \ldots$ and let $\mathcal{F}_n$ be the pre-task history, including current skill θ_n. Workflow choice $A_n = A^*(\theta_n; \theta_a)$ is $\mathcal{F}_n$-measurable and precedes the current task events. Use the potential events from Appendix B.2, and write $R_n = E_n I_{D,n}$ for the potential repair indicator under review. Its conditional probability is

$$\mathbb{P}(R_n = 1 \mid \mathcal{F}_n) = q(\theta_n; \theta_a).$$

This is a conditional event law at each task, not a requirement that outcomes be unconditionally independent across a skill-dependent history. Only V uses R_n in the update.

Under M, production supplies generative learning $-\ell'_M(\theta)$ and self-checking supplies practice $\kappa(1-\theta)$. Under P, neither activity occurs, leaving non-use decay $-\Delta\theta$. Under V, inspection supplies practice on every task, while generative learning replaces non-use decay only when $R_n = 1$. A repair provides the same per-event generative signal as manual production; this learning convention is separate from equality of their mean output losses. Beyond verification practice, passive exposure to correct AI outputs supplies no additional learning in the baseline; Appendix E.2 relaxes that restriction.

These signals are encoded by local task potentials

$$\begin{aligned}
\Phi_M(\theta) &= \ell_M(\theta) + \tfrac{\kappa}{2}(1-\theta)^2, \\
\Phi_V(\theta; R) &= R\ell_M(\theta) + (1-R)\tfrac{\Delta}{2}\theta^2 + \tfrac{\kappa}{2}(1-\theta)^2, \\
\Phi_P(\theta) &= \tfrac{\Delta}{2}\theta^2.
\end{aligned}$$

Differentiation holds the realized event R fixed and gives

$$G_A(\theta, R) = \begin{cases} \ell'_M(\theta) - \kappa(1-\theta), & A = M, \\ R\ell'_M(\theta) + (1-R)\Delta\theta - \kappa(1-\theta), & A = V, \\ \Delta\theta, & A = P. \end{cases}$$

The subscripted G_A is a potential-gradient signal, distinct from the utility gap G in Section 3. Conditional on a realized learning event, the baseline signal is deterministic; the repair event still makes the task-level verified update stochastic.

A bounded discrete update. Following Amari (1998) and Huang & Vishnoi (2026), use

$$F(\theta) = \frac{1}{\theta(1-\theta)}, \qquad 0 < \theta < 1.$$

This is the Fisher metric of a Bernoulli parameter, used as a geometry for bounded latent skill, not as a claim that skill itself is a Bernoulli outcome. A negative natural-gradient step is

$$\theta_{n+1} = \theta_n - \eta\theta_n(1-\theta_n)G_{A_n}(\theta_n, R_n). \tag{6}$$

Continuous differentiability of ℓ_M on $[0,1]$ implies $L := \|\ell'_M\|_\infty < \infty$. Since $\ell'_M \le 0$,

$$-(L+\kappa) \le G_A(\theta, R) \le \Delta.$$

Put $K_s := \max\{L+\kappa, \Delta\}$. A sufficient uniform condition is

$$\eta K_s \le 1. \tag{7}$$

Indeed, for any $x \in [0,1]$ and any realized signal,

$$\begin{aligned} 0 &\le x - \eta K_s x(1-x) \\ &\le x - \eta x(1-x)G_A(x, R) \\ &\le x + \eta K_s x(1-x) \le 1. \end{aligned}$$

The first inequality follows from $x[1-\eta K_s(1-x)] \ge 0$; the last follows from $(1-x)[1-\eta K_s x] \ge 0$. Both skill endpoints are invariant. A projected implementation can be defined for other step sizes by applying $\Pi_{[0,1]}$ to the proposed next state, but the exact expectation identity below is asserted only when projection is inactive. Projection is not an algebraically equivalent operation on conditional means.

Conditional mean learning and skill drift. For a candidate workflow, condition on $\mathcal{F}_n$; all skill-dependent coefficients below are then fixed. The manual and pure signals give

$$\begin{aligned} \mathbb{E}[G_M(\theta_n, R_n) \mid \mathcal{F}_n] &= \mathbb{E}[\ell'_M(\theta_n) - \kappa(1-\theta_n) \mid \mathcal{F}_n] \\ &= \ell'_M(\theta_n) - \kappa(1-\theta_n), \\ \mathbb{E}[G_P(\theta_n, R_n) \mid \mathcal{F}_n] &= \mathbb{E}[\Delta\theta_n \mid \mathcal{F}_n] = \Delta\theta_n. \end{aligned}$$

Writing $q_n = q(\theta_n; \theta_a)$, the verified signal satisfies

$$\begin{aligned} &\mathbb{E}[G_V(\theta_n, R_n) \mid \mathcal{F}_n] \\ &= \mathbb{E}[R_n\ell'_M(\theta_n) + (1-R_n)\Delta\theta_n - \kappa(1-\theta_n) \mid \mathcal{F}_n] \\ &= \ell'_M(\theta_n)\mathbb{E}[R_n \mid \mathcal{F}_n] + \Delta\theta_n\mathbb{E}[1-R_n \mid \mathcal{F}_n] - \kappa(1-\theta_n) \\ &= q_n\ell'_M(\theta_n) + (1-q_n)\Delta\theta_n - \kappa(1-\theta_n). \end{aligned}$$

Define the mean branch signal $\bar{G}_A(\theta; \theta_a)$ by these three expressions and set

$$f_A(\theta; \theta_a) := -\theta(1-\theta)\bar{G}_A(\theta; \theta_a).$$

Because A_n is selected before the current task events, the candidate identities apply to the selected workflow as well. Under (7),

$$\begin{aligned} \mathbb{E}[\theta_{n+1} - \theta_n \mid \mathcal{F}_n] &= \mathbb{E}[-\eta\theta_n(1-\theta_n)G_{A_n}(\theta_n, R_n) \mid \mathcal{F}_n] \\ &= -\eta\theta_n(1-\theta_n)\mathbb{E}[G_{A_n}(\theta_n, R_n) \mid \mathcal{F}_n] \\ &= -\eta\theta_n(1-\theta_n)\bar{G}_{A_n}(\theta_n; \theta_a) \\ &= \eta f_{A_n}(\theta_n; \theta_a). \end{aligned}$$

These are exactly the three branch drifts in (5). There is no q' term: the rule differentiates the task potential with its event fixed and then averages the signal under the current-state event law. Differentiating an expectation while also changing that law would add a response of the event distribution, which is not part of the learning rule. A mean skill potential, if desired, is an antiderivative of $\bar{G}_A$; it need not equal the expectation of the local task potential.

Three distinct levels of dynamics. Replacing each realized increment by its conditional mean defines a new, deterministic recursion

$$x^{\eta}_{n+1} = x^{\eta}_n + \eta f_{A^*(x^{\eta}_n;\theta_a)}(x^{\eta}_n;\theta_a), \qquad x^{\eta}_0 = \theta_0. \tag{8}$$

On the time scale $t = n\eta$, this is an Euler discretization of the selected branch wherever the workflow is constant. The interpretation as a mean learning drift follows the stochastic-approximation viewpoint (Benaïm, 1999; Huang & Vishnoi, 2026), but the following objects must be kept separate:

Object	Construction	Use in this paper
Realized task process	Actual event-dependent updates in (6).	Finite-task first passage in Appendix E.3.
Conditional-mean recursion	Replace each random increment by its conditional expectation, as in (8).	Defines the deterministic approximation.
Vanishing-step trajectory	A locally uniform limit of interpolations of the mean recursion as $\eta \downarrow 0$.	Convergence and path comparisons in Appendix C.

In general, $x^{\eta}_n \neq \mathbb{E}[\theta_n]$: $\mathbb{E}[f^\star(\theta_n)]$ need not equal $f^\star(\mathbb{E}[\theta_n])$. Nor does finite-time mean-field tracking by itself establish almost-sure convergence of a fixed-step random history.

Appendix C.2 defines the vanishing-step dynamics at utility ties and proves boundary trapping and convergence under finite regularity. Appendix C.5 treats degenerate cases; Appendix E.3 studies stochastic escape at fixed step size.

B.4 Connection to the static delegation and verification model

We identify a fixed-state correspondence with Huang et al. (2026) under zero–one loss and the cost restrictions below. For this subsection only, write the earlier model's delegation probability as $x \in [0,1]$ and its verification effort as $s \in [0,1]$, reserving d for our cost exponent. The remaining symbols $p_w, p_a, \phi, C_w, C_a, C_v, b_W, \ell_W$ below also belong only to the earlier model and its mapping. Here p_w, p_a are worker and AI success probabilities, $\phi(s)$ is detection probability conditional on an AI error, and $C_w, C_a, C_v(s)$ are execution and verification costs, with $\phi(0) = C_v(0) = 0$. The task-success probability and expected cost are

$$\begin{aligned} p_{\mathrm{succ}}(x,s) &= (1-x)p_w + xp_a + x(1-p_a)\phi(s)p_w, \\ C_{\mathrm{old}}(x,s) &= (1-x)C_w + x\big[C_a + C_v(s) + (1-p_a)\phi(s)C_w\big]. \end{aligned}$$

With success benefit b_W and failure penalty ℓ_W, its worker utility is $U_W = b_W p_{\mathrm{succ}} - \ell_W(1 - p_{\mathrm{succ}}) - C_{\mathrm{old}}$. For $B_W = b_W + \ell_W > 0$, the positive affine normalization

$$\frac{U_W - b_W}{B_W} = -(1 - p_{\mathrm{succ}}) - \gamma C_{\mathrm{old}}, \qquad \gamma = \frac{1}{B_W},$$

preserves the maximizer and puts utility on the loss–cost scale of (1).

Because utility is affine in x for each fixed s, an optimum can be attained at $x \in \{0,1\}$; interior delegation choices can be optimal only at ties. Let $s^\dagger$ maximize utility conditional on delegation. The three workflow categories correspond to $(x,s) = (0,0)$ for M, $(1,0)$ for P, and $(1,s^\dagger)$ for V, with $V = P$ if $s^\dagger = 0$. Under zero–one loss, p_w equals manual-output reliability and $1 - p_a$ is the AI-error probability; every AI error has unit severity. The following identifications therefore match losses and costs at the fixed state:

$$\begin{aligned} &\varepsilon_a = 1 - p_a, \qquad \bar{\ell}_a = 1, \qquad r_m = p_w, \qquad r_v = \phi(s^\dagger), \\ &c_g = C_w, \qquad c_s = 0, \qquad k_a = C_a, \qquad c_v = C_v(s^\dagger), \quad \lambda = 1. \end{aligned}$$

In particular, verified delegation has success probability

$$1 - \ell_V = p_a + (1-p_a)\phi(s^\dagger)p_w = p_{\mathrm{succ}}(1, s^\dagger)$$

and expected cost

$$\mathrm{cost}_V = C_a + C_v(s^\dagger) + (1-p_a)\phi(s^\dagger)C_w = C_{\mathrm{old}}(1, s^\dagger).$$

The manual and pure-delegation alternatives follow immediately.

This correspondence is pointwise, not an equivalence of the models over evolving skill: optimized effort $s^\dagger$ may depend on worker and AI characteristics in ways that do not preserve our separability or shape assumptions. The present model also permits nonbinary AI-error severity, discounted repair effort, and a separate manual self-checking cost. Its dynamic addition is to let workflow-dependent practice change the worker skill that determines future reliability, cost, and workflow choice.

The payoff correspondence does not specify a learning law.

B.5 Connection to the adaptive AI-reliance model

The model of Huang & Vishnoi (2026) couples worker skill with a continuous delegation fraction $p(t)$, without distinguishing verified from pure delegation. To recover its skill equation, suppress V, set $\kappa = 0$, and use $\ell_M(\theta) = (1-\theta)^2$. A fraction $1-p$ of manual tasks and p of purely delegated tasks then gives

$$\dot{\theta} = (1-p)f_M + pf_P = 2\theta(1-\theta)\left[(1-p)(1-\theta) - \tfrac{\Delta}{2}p\theta\right].$$

With $\Delta_{\text{old}} = \Delta/2$ and $t_{\text{old}} = 2t$, this becomes

$$\tfrac{d\theta}{dt_{\text{old}}} = \theta(1-\theta)\big[(1-p)(1-\theta) - \Delta_{\text{old}}p\theta\big],$$

the normalized skill equation of Huang & Vishnoi (2026) with default skill $\theta_d = 0$. Holding $p = p_0$ constant gives that paper's fixed-delegation benchmark.

Adaptive delegation requires an additional behavioral law; it does not follow from the skill equation alone. For expected loss $L(\theta, p) = (1-p)\ell_M(\theta) + p\ell_P(\theta_a)$, the earlier model's multiplicative natural-gradient update gives

$$\tfrac{dp}{dt_{\text{old}}} = \kappa_p p(1-p)\big[\ell_M(\theta) - \ell_P(\theta_a)\big] = \kappa_p p(1-p)\big[(1-\theta)^2 - (1-\theta_a)^2\big],$$

where the last equality uses $\ell_P(\theta_a) = (1-\theta_a)^2$, and κ_p is the reliance-adjustment rate in rescaled time, not our verification-practice parameter κ. Thus the models share manual learning and non-use decay but close the decision problem differently: reliance adjusts gradually in the earlier two-state system, whereas (1) selects a discrete workflow from current utility.

What verification adds. Verified delegation introduces a learning branch absent from both preceding models: human generative practice occurs only with probability $q = \varepsilon_a r_v$, while verification practice occurs on every task. In the finite regular case of Appendix C.2, this branch supports an interior attractor only at a stable root in the strict region R_V; alternatively, opposing manual and verified drifts can sustain skill at an attracting $M \to V$ boundary. A nonempty verification region alone is insufficient: skill may decline while review remains optimal and eventually enter pure delegation. The additional question is therefore whether human review preserves the competence needed to make that review sustainable, not only whether delegation can erode skill.

These correspondences match specified payoffs or skill drifts, not the full dynamic models.

C Proofs and additional results from Section 3

We prove the workflow, convergence, collapse, and AI-quality results from Section 3, stating the assumptions for each. The proofs proceed in the order

$$\text{workflow partition} \longrightarrow \text{selected phase line} \longrightarrow \text{collapse and cross-quality comparisons.}$$

The workflow result uses the model's shape restrictions; the finite basin formula and collapse corollary also use Assumption C.3. Pathwise AI-quality comparisons additionally require branchwise local Lipschitz continuity and an initial state strict at both qualities.

All convergence statements concern vanishing-step limits of the *conditional-mean* recursion in Appendix B.3, followed by their long-time limits. They do not classify the fixed-step stochastic process. Appendices C.5 and D.1.2 treat degenerate cases and critical parameter values, respectively.

C.1 Proof of Theorem 3.1: optimal workflow choice

We first establish the pure-delegation cutoff, then compare manual work with verified delegation above it, including ties.

Fix θ_a and suppress it from the notation. Recall $G(\theta) := \max\{U_M(\theta), U_V(\theta)\} - U_P$, $D(\theta) := U_V(\theta) - U_M(\theta)$, and $\tau_P := \inf\{\theta \in (0,1) : G(\theta) > 0\}$, with $\inf \varnothing := 1$. The cutoff τ_P marks the end of the skill range in which pure delegation remains optimal, possibly tied with another workflow. The key property is that once M or V strictly improves on P, that workflow remains strictly better at every higher skill.

Lemma C.1 (Pure-delegation cutoff). *For $A \in \{M, V\}$ and $0 < \theta_1 < \theta_2 < 1$, $U_A(\theta_1) > U_P$ implies $U_A(\theta_2) > U_P$. Consequently, $\{\theta \in (0,1) : G(\theta) > 0\} = (\tau_P, 1)$.*

Proof: Write $H_{MP} := U_M - U_P$ and $H_{VP} := U_V - U_P$. Equations (2)–(4) give

$$H_{MP} = \varepsilon_a \bar{\ell}_a - \ell_M + \gamma(k_a - c_g - c_s),$$
$$H_{VP} = qB - \gamma c_v, \qquad B := \bar{\ell}_a - \ell_M - \gamma\lambda c_g.$$

Here B is the gain from repairing a detected AI error rather than retaining it, net of repair cost. Thus qB is the expected repair gain, whereas γc_v is the inspection cost incurred on every verified task. The shape restrictions give $H'_{MP} = -\ell'_M - \gamma(c'_g + c'_s) \geq 0$, so a strict manual-work advantage persists at higher skill.

For verification, higher detection alone need not help when $B < 0$: it triggers more repairs whose costs exceed their benefits. Hence H_{VP} need not be globally nondecreasing. However, $H_{VP}(\theta_1) > 0$ implies $q(\theta_1)B(\theta_1) > \gamma c_v(\theta_1) \geq 0$, hence $q(\theta_1) > 0$ and $B(\theta_1) > 0$. Since $q' = \varepsilon_a r'_v \geq 0$, $B' = -\ell'_M - \gamma\lambda c'_g \geq 0$, and $c'_v \leq 0$, every $\theta_2 > \theta_1$ satisfies

$$H_{VP}(\theta_2) = q(\theta_2)B(\theta_2) - \gamma c_v(\theta_2) \geq q(\theta_1)B(\theta_1) - \gamma c_v(\theta_1) > 0.$$

Thus each workflow's strict advantage over P persists, and $\{G > 0\} = \{H_{MP} > 0\} \cup \{H_{VP} > 0\}$ is an upper interval. Continuity makes it open in $(0,1)$, so its infimum gives $\{G > 0\} = (\tau_P, 1)$, including the empty and full-interior cases. □

Above τ_P, only the sign of D is needed: it is constant on each connected component of $\{\theta \in (\tau_P, 1) : D(\theta) \neq 0\}$ and selects the unique optimal workflow there.

Theorem C.2 (Formal statement of Theorem 3.1). *The strict action regions are $R_P = \{G < 0\}$, $R_M = \{G > 0, D < 0\}$, and $R_V = \{G > 0, D > 0\}$. The cutoff τ_P marks where pure delegation ceases to be optimal: below it, P remains a maximizer, possibly tied; above it, P is strictly dominated and*

$$A^*(\theta) = \begin{cases} V, & D(\theta) > 0, \\ M, & D(\theta) < 0. \end{cases}$$

At $G = 0$, P ties with the human-inclusive workflows attaining $\max\{U_M, U_V\}$; at $G > 0$ and $D = 0$, exactly M and V are optimal. All ties are resolved by the model's fixed tie-breaking rule, including isolated zeros, zero intervals, and $D \equiv 0$.

For the finite representation, suppose $\tau_P < 1$, every zero of D in $(\tau_P, 1)$ is isolated, and exactly $n < \infty$ of them change sign. Write these sign-changing zeros as $\tau^{(1)} < \cdots < \tau^{(n)}$, set $\tau^{(0)} := \tau_P$ and $\tau^{(n+1)} := 1$, and let σ be the sign of the nonzero values of D on $(\tau^{(0)}, \tau^{(1)})$. Then, for $\theta \in (\tau^{(j)}, \tau^{(j+1)})$ with $D(\theta) \neq 0$,

$$A^*(\theta) = \begin{cases} V, & \sigma(-1)^j > 0, \\ M, & \sigma(-1)^j < 0, \end{cases} \qquad j = 0, \ldots, n.$$

If $\tau_P = 1$, there is no upper region and no σ is needed.

Proof: The strict-region identities and tie descriptions follow by comparing the three utilities. Lemma C.1 gives $G \leq 0$ below τ_P and $G > 0$ above it. If $0 < \tau_P < 1$, continuity also gives $G(\tau_P) = 0$. On each connected component of $\{G > 0, D \neq 0\}$, continuity prevents D from changing sign, so its sign selects the unique maximizer. This proves the general characterization without restricting the zero set.

Under the finite-representation assumptions, put $I_j := (\tau^{(j)}, \tau^{(j+1)})$. Every compact subinterval of I_j contains finitely many zeros: otherwise their accumulation point would be a non-isolated zero by continuity. Opposite signs at two points of I_j would therefore require a sign-changing zero between them, contrary to its definition. Thus all nonzero values on I_j have one sign; such values exist because the zeros are isolated. This defines σ on I_0, even when non-sign-changing zeros accumulate at τ_P. Each $\tau^{(j)}$ reverses the sign, giving $\operatorname{sgn} D(\theta) = \sigma(-1)^j$ on $I_j \cap \{D \neq 0\}$. The general characterization then yields the finite formula. □

Appendix D gives a concrete example: when $\theta_a = 0.3$, the strict workflow sequence is $P \to V \to M \to V$. Thus the verified-delegation region can be disconnected even though pure delegation never becomes optimal again after losing to a human-inclusive workflow.

The workflow partition determines the active branches. We next use their learning drifts to find the attractors and attraction basins.

C.2 Proof of Theorem 3.2: skill convergence and attraction basins

There are two limits, taken in a fixed order. We first let the task-update step size $\eta \downarrow 0$, taking locally uniform subsequential limits of the interpolated recursion $\theta^\eta_{n+1} = \theta^\eta_n + \eta f^\star(\theta^\eta_n)$, initialized at θ_0. Here θ^η_n denotes the deterministic mean iterate x^η_n in (8), not a realized random skill history. We then study $t \to \infty$ along these *vanishing-step trajectories*. The finite regular case gives the main-text max–min formula; Appendix C.5 treats degenerate cases.

Define

$$S_V := \{z \in R_V : f_V(z) = 0,\ f'_V(z) < 0\}.$$

Let B_{MV} be the interior boundaries with M selected immediately to the left and V immediately to the right, and put

$$S_{MV} := \{\tau \in B_{MV} : f_M(\tau) > 0 > f_V(\tau)\}, \qquad S := S_V \cup S_{MV}.$$

An inactive root of f_V is not included in S_V.

For $n = \lfloor t/\eta \rfloor$ and $u = t/\eta - n$, the interpolation is

$$\theta^\eta(t) = (1-u)\theta^\eta_n + u\theta^\eta_{n+1}.$$

To include workflow boundaries, define the local drift envelopes

$$\underline{f}(x) := \lim_{\delta \downarrow 0} \inf_{\substack{y \in [0,1] \\ |y-x|<\delta}} f^\star(y), \qquad \overline{f}(x) := \lim_{\delta \downarrow 0} \sup_{\substack{y \in [0,1] \\ |y-x|<\delta}} f^\star(y).$$

Let $\mathcal{Z} := \{x \in [0,1] : \underline{f}(x) \le 0 \le \overline{f}(x)\}$. Both envelopes equal f_A on R_A. The set $\mathcal{Z}$ contains the endpoints and all possible limiting states, not only attractors.

Finite regular case. We use the following regularity conditions.

Assumption C.3 (Finite regularity of the selected dynamics)**.** *Fix AI quality. The following conditions hold:*

1. *The interior states with a nonunique optimal workflow form a finite set. At each such state τ, the unique workflows A_- and A_+ on its two sides are distinct, and the maximizing set at τ is exactly $\{A_-, A_+\}$.*
2. *$f_M(\theta) > 0$ for every $\theta \in (0,1)$.*
3. *The active verified roots are finite and hyperbolic: at every $z \in R_V$ with $f_V(z) = 0$, the derivative $f'_V(z)$ exists and is nonzero.*
4. *Both adjacent branch drifts are nonzero at every workflow tie.*

The assumption excludes isolated non-switching ties, tie intervals, triple ties, and branch roots at workflow boundaries. Either maximizing workflow may be selected at a genuine tie by a fixed rule. These conditions are additional to the shape restrictions in Section 2; Appendix D.1.2 treats exceptional parameter values.

Theorem C.4 (Formal statement of Theorem 3.2). *Under Assumption C.3, all vanishing-step trajectories from a strict initial state $\theta_0 \in R_P \cup R_M \cup R_V$ converge to the same limiting skill,*

$$\theta^\infty(\theta_0) = \begin{cases} \max\{z \in S \cup \{0\} : z \le \theta_0\}, & f^\star(\theta_0) < 0, \\ \theta_0, & f^\star(\theta_0) = 0, \\ \min\{z \in S \cup \{1\} : z \ge \theta_0\}, & f^\star(\theta_0) > 0. \end{cases}$$

The interior locally attracting skill levels are exactly $S = S_V \cup S_{MV}$. Initialization at $\tau \in S_{MV}$ gives $\theta(t) = \tau$ for all $t \ge 0$.

Proof: Existence and local dynamics. This step requires neither finite regularity nor a restriction on the fixed tie rule. Write $f_A(x) = x(1-x)g_A(x)$ and choose K_g bounding the continuous functions $|g_A|$. For $\eta K_g \le 1$,

$$0 \le x - \eta K_g x(1-x) \le x + \eta f^\star(x) \le x + \eta K_g x(1-x) \le 1.$$

Thus the interpolations are uniformly bounded and Lipschitz. Arzelà–Ascoli and diagonal extraction give locally uniform limits, which are absolutely continuous and remain in $[0,1]$. If a limiting path stays in a neighborhood N over $[s,t]$, the discrete slopes lie between the infimum and supremum of $f^\star$ on any slightly larger neighborhood. Passing to the limit, taking difference quotients at differentiability times, and shrinking the neighborhood gives

$$\dot\theta(t) \in [\underline{f}(\theta(t)), \overline{f}(\theta(t))] \quad \text{for almost every } t.$$

In particular, $\dot\theta = f_A(\theta)$ on any open interval where the selected workflow is constant and equal to A, including tie intervals. If $z = \lim_{t\to\infty}\theta(t)$ exists, then $z \in \mathcal{Z}$: otherwise a neighborhood of z has, for example, $f^\star \ge c > 0$, forcing $\theta(t) - \theta(T) \ge c(t-T)$ for large T, a contradiction. The negative case is identical.

Phase-line direction. Under finite regularity, the workflow ties and active verified roots divide $(0,1)$ into finitely many intervals with a nonzero drift of constant sign. A nonmonotone path would traverse one such interval in both directions, contradicting its local differential equation. Every limiting path is therefore monotone and convergent. A nonzero initial drift fixes its direction; a zero initial drift leaves every recursion stationary. The local possibilities are

Local object	Drift signs (left, right)	Dynamic role
Active stable V-root	$(+,-)$	Attractor in S_V
Active unstable V-root	$(-,+)$	Repels; may separate basins
Inward $M \to V$ switch	$(+,-)$	Attractor in S_{MV}
Outward workflow switch	$(-,+)$	Repels
Same-sign workflow switch	$(+,+)$ or $(-,-)$	Crossed in that direction

Since $f_M > 0 > f_P$ and Lemma C.1 places P below the other workflows, no boundary involving P is inward pointing. A root of f_V outside R_V is inactive and does not arrest the selected dynamics.

Trapping at an inward switch. For $\tau \in S_{MV}$, choose $\delta, K, m > 0$ with $m \le f_M \le K$ on $[\tau-\delta, \tau]$ and $-K \le f_V \le -m$ on $[\tau, \tau+\delta]$. For $K\eta < \delta$, the interval $J_\eta = [\tau - K\eta, \tau + K\eta]$ is forward invariant: every step points toward τ and overshoots by at most $K\eta$, including either tie choice at τ. Starting at x in this neighborhood, entry into J_η takes physical time at most $|x-\tau|/m + \eta$.

Suppose $\theta^{\eta_j} \to \theta$ and $\theta(T) = \tau$, allowing $T = 0$. Set $n_j = \lfloor T/\eta_j \rfloor$ and $d_j = |\theta^{\eta_j}_{n_j} - \tau| \to 0$. The remaining entry time is at most $d_j/m + \eta_j \to 0$; hence, for every $H > T + \epsilon > T$ and all sufficiently large j,

$$\sup_{s \in [T+\epsilon, H]} |\theta^{\eta_j}(s) - \tau| \le K\eta_j \longrightarrow 0.$$

Thus $\theta(t) = \tau$ for all $t \ge T$. Finite-step iterates may switch across the boundary, but their shrinking oscillations select a stationary vanishing-step limit.

Roots, passage, and the basin formula. At an active hyperbolic root z, differentiability gives $|f_V(x)| \le L|x-z|$ locally. For $\eta L < 1$, an Euler step cannot cross z; the derivative sign distinguishes attraction from repulsion. At a same-sign switch, both adjacent drifts and the tie-selected

drift are bounded away from zero with that sign, forcing passage. An outward switch cannot be approached from either side. Consequently, a nonstationary trajectory stops at the first element of S in its direction of motion, or at the corresponding endpoint if none intervenes: the inclusion excludes other stopping states, and the root bound or switching trap prevents passage through an attractor. This proves the max–min formula, without imposing finite regularity on separately analyzed critical parametric cases. □

Corollary C.5 (Manual-practice share at an attracting workflow boundary)**.** *Under Assumption C.3, let $\tau \in S_{MV}$ and consider the conditional-mean recursion after it has entered the invariant tube J_η constructed above. Over $N_\eta = \lfloor h/\eta \rfloor$ updates, where $h > 0$ is fixed, let α_η be the fraction of manual updates. Then*

$$\alpha_\eta \longrightarrow \alpha^\star(\tau) := \frac{-f_V(\tau)}{f_M(\tau) - f_V(\tau)} \in (0,1) \qquad \text{as } \eta \downarrow 0.$$

Proof: All iterates in the window lie within $K\eta$ of τ, and only M or V can be selected there. Telescoping the updates gives

$$\begin{aligned} \frac{\theta^\eta_{n+N_\eta} - \theta^\eta_n}{N_\eta \eta} &= \frac{1}{N_\eta} \sum_{j=n}^{n+N_\eta-1} f_{A^*(\theta^\eta_j)}(\theta^\eta_j) \\ &= \alpha_\eta f_M(\tau) + (1-\alpha_\eta) f_V(\tau) + o(1). \end{aligned}$$

Uniform continuity of the two branch drifts makes the last error tend to zero. The numerator on the left is $O(\eta)$ and $N_\eta \eta \to h > 0$, so the left side tends to zero. Solving the resulting balance gives the claimed limit. □

The boundary supports competence through an endogenous balance of practice and loss. The width of the manual-work region does not determine the fraction of tasks performed manually; that fraction is determined by the two local drifts. This share concerns the conditional-mean recursion and its vanishing-step limit, not the fixed-step stochastic escape experiment.

The next corollary rules out both interior sustaining mechanisms.

C.3 Proof of Corollary 3.3: conditions for global skill collapse

The collapse corollary uses Assumption C.3, in particular its nonzero boundary drifts.

Corollary C.6 (Restatement of Corollary 3.3). *Under the finite regularity conditions of Theorem C.4, if $R_M = \varnothing$ and $f_V(\theta) < 0$ for every $\theta \in R_V$, then every initial skill $\theta_0 \in (0,1)$ has the vanishing-step long-run limit $\theta^\infty(\theta_0) = 0$.*

Proof: The genuine two-workflow-switch condition excludes manual work even at ties: a tie involving M would require a neighboring strict manual region. Thus only P and V can be selected in the interior. At any P/V boundary τ, continuity gives $f_V(\tau) \le 0$ from the verified side, and the nonzero adjacent-drift condition makes this inequality strict. Since $f_P(\tau) < 0$ as well, either permitted tie choice has negative drift. In particular, such a boundary cannot arrest decline.

Every interior state consequently has a neighborhood on which the selected drift is bounded above by a strictly negative constant; equivalently, $\overline{f}(x) < 0$ for every $x \in (0,1)$ in the notation of Appendix C.2. The finite-case convergence result gives a decreasing trajectory from each strict initial state. Its limit is below 1 and belongs to $\mathcal{Z}$, which contains no interior point under these drift bounds. Hence every subsequential trajectory has the same long-run value,

$$\theta^\infty(\theta_0) = 0 \qquad \text{for every } \theta_0 \in R_P \cup R_M \cup R_V.$$

At an initial workflow tie, both adjacent drifts are strictly negative; the local drift bound forces immediate passage into the lower strict region. The same limit therefore holds for every $\theta_0 \in (0,1)$. If both P and V regions are nonempty, Lemma C.1 places P below V. A worker starting in R_V therefore loses skill while verifying, passes through the positive P/V boundary, and then converges to zero under pure delegation. If $\tau_P = 0$, collapse need not involve a switch into P. No finite-time arrival at zero is asserted. □

Why the boundary condition matters. Negativity of f_V on the *strict* verified region alone is not enough at a degenerate boundary. If $f_V(\tau_P) = 0$ and verified trajectories approach that boundary only asymptotically, they need not enter P. Appendix D.1.2 exhibits exactly this case at a_5. The regularity condition excludes this critical equality.

C.4 Proof of Theorem 3.4: effects of AI quality

Pathwise and long-run comparisons require additional regularity beyond the workflow and local equilibrium comparisons. Write $G(\theta; \theta_a) := \max\{U_M(\theta), U_V(\theta; \theta_a)\} - U_P(\theta_a)$.

Theorem C.7 (Formal statement of Theorem 3.4). *Fix $\theta_a' > \theta_a$, holding all other primitives fixed. Then $\tau_P(\theta_a') \geq \tau_P(\theta_a)$, $\{G(\cdot; \theta_a) \leq 0\} \subseteq \{G(\cdot; \theta_a') \leq 0\}$, $R_M(\theta_a') \subseteq R_M(\theta_a)$, and $f_V(\theta; \theta_a') \leq f_V(\theta; \theta_a)$ for every $\theta \in (0, 1)$. An active verified root $z_V(\theta_a)$ satisfies $z_V'(\theta_a) \leq 0$ wherever f_V is continuously differentiable near the branch and $\partial_\theta f_V(z_V(\theta_a); \theta_a) < 0$.*

If both qualities satisfy the finite regularity conditions of Theorem C.4, and their branch drifts are locally Lipschitz on $(0, 1)$, then $R_P(\theta_a) \subseteq R_P(\theta_a')$. For a common initial state θ_0 strict at both qualities, the unique vanishing-step paths and their long-run limits satisfy

$$\theta(t; \theta_0, \theta_a') \leq \theta(t; \theta_0, \theta_a) \quad (t \geq 0), \qquad \theta^\infty(\theta_0; \theta_a') \leq \theta^\infty(\theta_0; \theta_a).$$

Proof: We first order workflow choices and learning drifts, then compare equilibria and the time needed to reach each skill level.

Workflow choice and learning. For the verified–pure comparison, write

$$H_{VP} := U_V - U_P = qB - \gamma c_v, \qquad B := \bar{\ell}_a - \ell_M - \gamma\lambda c_g.$$

If $H_{VP}(\theta; \theta_a') > 0$, then $B(\theta; \theta_a') > 0$. Since $q(\theta; \theta_a) \geq q(\theta; \theta_a')$ and $B(\theta; \theta_a) \geq B(\theta; \theta_a')$, it follows that $H_{VP}(\theta; \theta_a) \geq H_{VP}(\theta; \theta_a') > 0$. Likewise, $U_M - U_P$ weakly decreases with AI quality because U_M is unchanged and U_P weakly increases. Thus $\{G(\cdot; \theta_a') > 0\} \subseteq \{G(\cdot; \theta_a) > 0\}$, implying $\tau_P(\theta_a') \geq \tau_P(\theta_a)$ by Lemma C.1. Equivalently, the weak-optimality set $\{G \leq 0\}$ for P expands. This does not yet imply nesting of the strict region $R_P = \{G < 0\}$: an old strict inequality can become a tie.

Both AI-mediated utilities weakly increase with quality. For V, write

$$U_V = -\gamma(k_a + c_v) - \varepsilon_a\big[(1 - r_v)\bar{\ell}_a + r_v\ell_M + \gamma\lambda r_v c_g\big].$$

Here all terms in brackets are nonnegative, while ε_a and $\bar{\ell}_a$ weakly decrease. Consequently, strict optimality of M at θ_a' implies its strict optimality at θ_a, proving $R_M(\theta_a') \subseteq R_M(\theta_a)$.

For interior skill, let $\Lambda(\theta) := \Delta\theta - \ell_M'(\theta) > 0$. The branch differences are

$$f_M - f_V = \theta(1 - \theta)(1 - q)\Lambda \geq 0,$$
$$f_V - f_P = \theta(1 - \theta)[q\Lambda + \kappa(1 - \theta)] \geq 0.$$

Moreover, $\partial_q f_V = \theta(1 - \theta)\Lambda > 0$. Since AI improvement weakly lowers q, it follows that $f_V(\theta; \theta_a') \leq f_V(\theta; \theta_a)$.

Active verified equilibria. Let $z_V(\theta_a)$ remain in $R_V(\theta_a)$, with $f_V(z_V(\theta_a); \theta_a) = 0$. Under the stated differentiability and stability conditions, the implicit-function theorem at the active root gives

$$z_V'(\theta_a) = -\frac{\partial_{\theta_a} f_V}{\partial_\theta f_V} \leq 0, \qquad \partial_{\theta_a} f_V = \theta(1 - \theta)\varepsilon_a'(\theta_a) r_v(\theta)\Lambda(\theta) \leq 0.$$

This comparison applies while the root remains active and stable; it does not compare different attractors.

Conditional finite regular comparison. Now impose the finite regularity and local Lipschitz conditions. Any $\theta \in R_P(\theta_a)$ has a neighborhood on which $G(\cdot; \theta_a) < 0$, hence $G(\cdot; \theta_a') \leq 0$. If $G(\theta; \theta_a') = 0$, the genuine two-workflow switch at that tie would give $G(\cdot; \theta_a') > 0$ on one side, a contradiction. Thus $R_P(\theta_a) \subseteq R_P(\theta_a')$.

For brevity write $F_a(x) := f^\star(x;\theta_a)$ and $F_b(x) := f^\star(x;\theta'_a)$. Away from the finite tie sets, the utility comparisons permit only $M \to V$, $M \to P$, or $V \to P$ as quality increases. Together with the branch inequalities, this gives

$$F_b(x) \le F_a(x) \quad \text{away from the two finite tie sets.}$$

This pointwise comparison alone does not invoke a smooth-ODE comparison theorem: the selected drifts can jump at workflow boundaries. We instead compare travel times along their monotone phase lines.

First, the vanishing-step path from a strict initial state is unique at each quality. Inside a strict region, local Lipschitz continuity gives a unique branch solution; a branch root cannot be reached in finite time from elsewhere. At a same-sign switch, $|F| \ge m > 0$ locally, so passage through $[\tau-\epsilon, \tau+\epsilon]$ takes at most $2\epsilon/m$: there is no waiting at the switch. An inward switch instead selects permanent sticking by the $O(\eta)$ trapping estimate in Appendix C.2; repelling switches cannot be reached from a strict initial state. Finally, $|f_A(x)| \le Kx(1-x)$ for a common K implies $\dot\theta \ge -K\theta$ and $\dot\theta \le K(1-\theta)$. Integration gives

$$\theta_0 e^{-Kt} \le \theta(t) \le 1 - (1-\theta_0)e^{-Kt},$$

so endpoints are not reached in finite time. These rules uniquely join the finitely many branch segments. Consequently, all interpolations converge locally uniformly to that path, and Theorem C.4 makes it monotone.

Let $x_a(t)$ and $x_b(t)$ denote the two paths from the common strict initial state θ_0, and fix $t > 0$. If $F_b(\theta_0) \le 0 \le F_a(\theta_0)$, monotonicity gives $x_b(t) \le \theta_0 \le x_a(t)$, including stationary initial roots. If both initial drifts are positive, choose any $\theta_0 < x < x_b(t)$. On (θ_0, x), $F_a \ge F_b > 0$ off the ties. This also rules out a stopping point for x_a before x: an active hyperbolic root or inward switch would require a negative drift on one adjacent side. Separating variables on each strict region and adding the travel times across the same-sign switches therefore gives

$$T_a(x) = \int_{\theta_0}^{x} \frac{du}{F_a(u)} \le \int_{\theta_0}^{x} \frac{du}{F_b(u)} = T_b(x) < t.$$

The finitely many tie values do not affect these integrals. Hence $x_a(t) \ge x$; letting $x \uparrow x_b(t)$ gives $x_a(t) \ge x_b(t)$, even after x_b sticks at a boundary.

If both initial drifts are negative, choose $x_a(t) < x < \theta_0$. Now $F_b \le F_a < 0$ along this interval, so the same argument, with downward travel times, yields

$$T_b(x) = \int_{x}^{\theta_0} \frac{du}{-F_b(u)} \le \int_{x}^{\theta_0} \frac{du}{-F_a(u)} = T_a(x) < t.$$

Thus $x_b(t) \le x$; letting $x \downarrow x_a(t)$ proves the remaining case. This establishes $x_b(t) \le x_a(t)$ for every $t \ge 0$. Taking the long-run limit only after the vanishing-step limit gives

$$\theta^\infty(\theta_0;\theta'_a) = \lim_{t\to\infty} x_b(t) \le \lim_{t\to\infty} x_a(t) = \theta^\infty(\theta_0;\theta_a).$$

□

Scope of the comparison. The pathwise and long-run orderings assume finite regularity, locally Lipschitz branch drifts, and an initial state strict at both AI qualities. They do not cover arbitrary initialization at workflow ties or the passive-learning extension in Appendix E.2.

C.5 Degenerate cases and general vanishing-step limits

Without Assumption C.3, compactness and a necessary differential inclusion survive, but the finite max–min basin formula and uniqueness of long-run skill need not. The envelopes $\underline{f}, \overline{f}$ and the possible limiting-state set $\mathcal{Z}$ are those defined in Appendix C.2.

Compactness holds for any fixed tie rule. To establish monotonicity here, we require a fixed priority ordering at ties, such as $P \succ V \succ M$. This extra condition is not needed for the finite regular theorem.

Theorem C.8 (**Subsequential vanishing-step limits).** *For any fixed tie rule, every sequence $\eta \downarrow 0$ has a subsequence whose interpolations converge locally uniformly to an absolutely continuous trajectory in $[0,1]$ satisfying $\dot{\theta} \in [\underline{f}(\theta), \overline{f}(\theta)]$ almost everywhere. Any existing long-run limit lies in $\mathcal{Z}$. Under a fixed priority ordering, every such trajectory from $\theta_0 \in R_M \cup R_V \cup R_P$ is monotone and satisfies*

$$\lim_{t\to\infty} \theta(t) \in \begin{cases} \mathcal{Z} \cap [0, \theta_0), & f^\star(\theta_0) < 0, \\ \{\theta_0\}, & f^\star(\theta_0) = 0, \\ \mathcal{Z} \cap (\theta_0, 1], & f^\star(\theta_0) > 0. \end{cases}$$

Different subsequences need not select the same path or limiting skill.

Proof: The compactness, inclusion, and limiting-state arguments in Appendix C.2 apply without finite regularity. It remains to prove monotonicity under fixed priority. For each pair A, B, the equality set $E_{AB} = \{x \in (0,1) : U_A(x) = U_B(x)\}$ is relatively closed, so its boundary is nowhere dense. Every open skill interval therefore contains a smaller open interval avoiding the boundaries of all three equality sets. On a connected subinterval, each utility difference is either identically zero or has fixed strict sign. The maximizing set is constant, and fixed priority selects one workflow A throughout it.

A nonmonotone trajectory would traverse some open skill interval both upward and downward. Within it choose a constant-workflow subinterval as above. If f_A is nonzero somewhere, continuity gives a smaller interval with $f_A \geq c > 0$ or $f_A \leq -c < 0$, permitting passage in only one direction. If f_A vanishes throughout, passage is impossible. Both cases contradict the two traversals. Thus the path is monotone and bounded. At a strict initial state, nonzero drift gives initial motion in its sign direction, making the stated limiting inequality strict; zero drift makes every initialized recursion constant. □

The envelope inclusion is a necessary property of vanishing-step limits; it does not say that every solution of the inclusion is selected by the recursion. Accordingly, $\theta^\infty(\theta_0)$ denotes a single limiting skill only where the selected trajectories have long-run limits that agree. The concrete critical cases in Appendix D.1.2 are resolved directly from the specialized drifts and the stated tie rule.

D Parametric illustrations and counterfactual comparisons

The general results of Appendix C do not depend on the parameter choices here. We first analyze the baseline workflow regions, learning equilibria, and critical AI qualities. We then vary workflow incentives with learning fixed and compare the menus with and without verification. These are illustrations, not empirical estimates.

Construction	What is fixed	What varies
Baseline figures (Appendix D.1)	The baseline functions and scalar parameters below.	AI quality θ_a.
Verification-option comparison (Appendix D.4)	The same primitives for both menus, with $\gamma = 0.32$ and baseline c_s.	The menu $\{M, P\}$ versus $\{M, V, P\}$; the proposition fixes $\theta_a = 1/2$, while its additional plots vary θ_a.
Common-drift comparison (Appendix D.2)	$\theta_a = 1/2$, $\lambda = 0.7$, $c_s = 0.04 + 0.05(1-\theta)$, all other utility primitives, and all three learning drifts.	Only cost sensitivity γ.

D.1 Baseline specification and parametric analysis

We first derive the workflow regions and then locate the learning equilibria within them.

D.1.1 Primitives and workflow regions

Set $\rho = 1$ and $p_m = p_v = 2$ in Section 2, so $r_m(\theta) = r_v(\theta) = 2\theta - \theta^2$ and $\ell_M(\theta) = (1-\theta)^2$, following the squared-loss baseline of Huang & Vishnoi (2026). Reliability is strictly concave and reaches one at full skill. For verification, it is also the probability of detecting an error in at least

one of two independent inspection opportunities, each successful with probability θ. Using the same curve removes a reliability difference without equating production and verification or their learning effects. We set $\varepsilon_a(\theta_a) = \bar{\ell}_a(\theta_a) = 1 - \theta_a$, hence $\ell_P(\theta_a) = (1 - \theta_a)^2$; equal improvement in the two AI-error margins is a simplifying choice, not an empirical restriction.

The affine costs ($d = 1$) and scalar parameters are

$$c_g(\theta) = c_v(\theta) = 0.1 + 0.4(1 - \theta), \qquad c_s(\theta) = 0.04 + 0.1(1 - \theta),$$
$$k_a = 0, \qquad \lambda = 0.5, \qquad \gamma = 0.3, \qquad \kappa = 1.25, \qquad \Delta = 3.$$

Positive floors retain effort at full skill, while declining costs capture greater efficiency. Self-checking is cheaper than generation or AI-output inspection; $c_g = c_v$ gives inspection no separate cost advantage. As in the negligible-AI-cost benchmark of Huang et al. (2026), $k_a = 0$; $\lambda = 0.5$ allows partial draft reuse without changing repair quality or learning exposure. Weighted manual cost ranges from 0.042 to 0.192 on the unit loss scale. Verification practice, $1.25(1 - \theta)$, is weaker per exposure than the generative force $2(1 - \theta)$, but its balance with decay also depends on skill and repair frequency; Appendix E.1 varies these rates.

Utility comparisons. Write $x = 1 - \theta$ and $e = 1 - \theta_a$. The utilities specialize to

$$U_M = -x^2 - 0.15x - 0.042, \qquad U_P = -e^2,$$
$$U_V = -ex^2(e + 1 - x^2) - 0.3(0.1 + 0.4x)\left[1 + 0.5e(1 - x^2)\right].$$

The gains over pure delegation and the verified–manual gap are therefore

$$\begin{aligned} H_{MP} &:= U_M - U_P = e^2 - x^2 - 0.15x - 0.042, \\ H_{VP} &:= U_V - U_P = e(1 - x^2)\left[e - x^2 - 0.15(0.1 + 0.4x)\right] - 0.3(0.1 + 0.4x), \\ D &:= U_V - U_M = H_{VP} - H_{MP}. \end{aligned} \tag{9}$$

For $A \in \{M, V\}$, let $\tau_{AP} := \inf\{\theta \in (0, 1) : H_{AP}(\theta) > 0\}$, with $\inf \varnothing := 1$. By Lemma C.1, each positive-gain set is an upper interval, so $\tau_P = \min\{\tau_{MP}, \tau_{VP}\}$. Above τ_P, the sign of D selects V or M; roots below it are dominated ties, not workflow boundaries. Throughout this illustration the fixed tie priority is $P \succ V \succ M$. The finite-task experiment states its own convention in Appendix E.3; its results are not reinterpreted under this rule.

At most two manual–verified crossings. In the variable x, the gap is

$$\widetilde{D}(x; e) = ex^4 + 0.06ex^3 + (1 - e^2 - 0.985e)x^2 + (0.03 - 0.06e)x + 0.012 - 0.015e.$$

For $e \le 1/2$, all coefficients are nonnegative and the constant is positive, so $\widetilde{D} > 0$. For $e > 1/2$, $\widetilde{D}_x(0) < 0$, $\widetilde{D}_x(1) = 2.03 + 2.15e - 2e^2 > 0$, and $\widetilde{D}_{xxx} = 24ex + 0.36e > 0$ on $[0, 1]$. Thus $\widetilde{D}_x$ is strictly convex and has exactly one zero: $\widetilde{D}$ first decreases and then increases, allowing at most two roots. In the skill coordinate θ, denote the lower and upper roots, when present, by $\tau_M^- < \tau_M^+$; at their collision use their common value. Only roots above τ_P are active, and these are the $\tau^{(j)}$ of Theorem 3.1.

At $\theta_a = 0.3$, the active boundaries in Figure 1a are approximately $(\tau_P, \tau_M^-, \tau_M^+) = (0.370, 0.498, 0.935)$, giving $P \to V \to M \to V$ as skill increases. At $\theta_a = 0.356$, they are $(0.407, 0.775, 0.803)$: manual work survives only in a narrow interval. These values solve the utility equalities with the third workflow checked, rather than being read from plotted pixels.

D.1.2 Skill dynamics and critical AI qualities

Substituting $q = (1 - \theta_a)(2\theta - \theta^2)$ into (5) gives

$$\dot{\theta} = \begin{cases} \frac{13}{4}\theta(1 - \theta)^2, & A^* = M, \\ \theta(1 - \theta)\left[\frac{5}{4} - \frac{1 + 16\theta_a}{4}\theta - (1 - \theta_a)\theta^3\right], & A^* = V, \\ -3\theta^2(1 - \theta), & A^* = P. \end{cases} \tag{10}$$

The verified bracket is $(2q + \kappa)(1 - \theta) - \Delta(1 - q)\theta$; its quadratic terms cancel, leaving $H(\theta; \theta_a) := \frac{5}{4} - \frac{1 + 16\theta_a}{4}\theta - (1 - \theta_a)\theta^3$. Since $\partial_\theta H = -\frac{1 + 16\theta_a}{4} - 3(1 - \theta_a)\theta^2 < 0$, $H(0) = 5/4$, and

$H(1) = -3\theta_a$, there is exactly one interior verified root $\theta_V^\dagger$ for $\theta_a > 0$; at $\theta_a = 0$ it is the endpoint 1. The verified drift points toward this root.

This single-root branch does not determine the selected dynamics. Its root is an active equilibrium only in R_V. If an active upper boundary τ_M^+ instead satisfies $H(\tau_M^+; \theta_a) < 0$, then $f_M(\tau_M^+) > 0 > f_V(\tau_M^+)$ creates a switching-supported limit. Thus multiple attractors arise from branch activation and workflow switching, not from multiple verified-learning equilibria. Away from critical ties, the possible limits are 0, an active $\theta_V^\dagger$, an inward-pointing τ_M^+, and 1. A strict initial state follows its selected drift to the first attracting limit in that direction; a zero drift leaves skill unchanged.

Critical values and attraction basins. With at most two M/V crossings and one verified root, the ordering changes only when a workflow boundary reaches an endpoint, meets the P cutoff, or collides with another boundary, or when the verified root meets an active boundary. The six feasible events below distinguish workflow-geometry changes from basin changes; the polynomial checks exclude additional events in $\theta_a \in (0, 1)$. Their defining equations, not the rounded values, specify the critical cases.

Cutoff	Defining condition	Approx. value	Type and effect
a_1	$D(1; a_1) = 0$	0.200	Geometry/basin: upper V appears; the limit leaves 1.
a_2	$H_{MP} = H_{VP} = 0$	0.227	Geometry: a triple tie creates the lower V interval.
a_3	$H(\tau_M^-; a_3) = 0$	0.342	Basin: the verified root enters the lower V interval.
a_4	$D = \partial_\theta D = 0$	0.356	Geometry/basin: M and its switching support disappear.
a_5	$H_{VP}(\theta_V^\dagger(a_5); a_5) = 0$	0.502	Basin: the verified root meets τ_P; collapse follows above a_5.
a_6	$H_{VP}(1; a_6) = 0$	0.819	Geometry: V disappears from the interior.

At a_2 and a_4, both equations hold at the same skill in $(0, 1)$. For a_3, require the lower root ($\partial_\theta D < 0$); at a_3 and a_4, the common M/V utility must exceed U_P. These eligibility checks exclude dominated crossings and irrelevant polynomial roots. Using rational polynomial coefficients, elimination of x by resultants and isolation of all real parameter roots in $e \in (0, 1)$ to width at most 10^{-12} yield one feasible common skill for each of a_2–a_5. Two additional parameter candidates for a_4 have no common skill in $(0, 1)$. Substitution into the original equations gives residuals below 10^{-8}.

The resulting phase sequence is summarized below. For interior initial skills, $\theta_0 \le \tau_P$ selects P and converges to zero; the last column therefore concerns only interior initial skills above τ_P.

AI quality	Workflows as skill increases	Limit for $\theta_0 > \tau_P$
$[0, a_1)$	$P \to M$	1
(a_1, a_2)	$P \to M \to V$	τ_M^+
(a_2, a_3)	$P \to V \to M \to V$	τ_M^+
(a_3, a_4)	$P \to V \to M \to V$	$\theta_V^\dagger$ if $\theta_0 \le \tau_M^-$; τ_M^+ if $\theta_0 > \tau_M^-$
(a_4, a_5)	$P \to V$	$\theta_V^\dagger$
(a_5, a_6)	$P \to V$	0
$(a_6, 1]$	P	No initial skill lies above $\tau_P = 1$.

In particular, for $a_4 < \theta_a < a_5$, the verified root lies strictly above τ_P. Trajectories initialized above the cutoff stay in V and approach that root asymptotically, without switching to P.

What the basin plot adds. Figure 1b shows how a small manual region can support a much larger basin. At $\theta_a = 0.356$, its width is only $\tau_M^+ - \tau_M^- \approx 0.028$, but the high-skill basin has width $1 - \tau_M^- \approx 0.225$. As $\theta_a \uparrow a_4$, the manual interval shrinks to $\theta \approx 0.790$; when it disappears, workers in its former high-skill basin instead converge to $\theta_V^\dagger(a_4) \approx 0.644$, a drop of about 0.146. Losing a small region of incentives for manual practice can therefore remove substantial long-run skill support. This does not mean that a small fraction of manual tasks suffices: at $\theta_a = 0.356$,

Corollary C.5 gives $\alpha^\star(\tau_M^+) \approx 0.40$ at $\tau_M^+ \approx 0.803$. Manual-region width, basin width, and local manual-task share are distinct.

After manual work disappears, both the positive basin and its target shrink: at $\theta_a = 0.4$ and 0.5, $(\tau_P, \theta_V^\dagger) \approx (0.439, 0.604)$ and $(0.521, 0.524)$, respectively. Above a_5, all interior skills collapse even though V remains optimal on an upper interval until a_6. Thus the loss of sustainable capability precedes the disappearance of verification from the workflow map.

Critical equalities and endpoints. At collisions and tangencies, the regular-case formula need not apply; use the specialized drifts and fixed tie rule directly:

- At a_1, high-skill interior trajectories still approach 1. At a_2, the triple tie selects P, while the adjoining positive basin still approaches τ_M^+.
- At a_3, $\theta_V^\dagger = \tau_M^-$: the lower verified interval approaches this point from below, initialization there stays fixed, and skills above it approach τ_M^+.
- At a_4, the manual interval vanishes at $\theta \approx 0.790$. The tie selects V, so its former high-skill basin joins the verified-root basin.
- At a_5, $\theta_V^\dagger = \tau_P \approx 0.523$ is a boundary root, not an active verified equilibrium: the tie selects P with negative drift. For $\theta_0 > \tau_P$, however, $f_V'(\tau_P) < 0$ implies $\int_{\tau_P}^{\theta_0} [-f_V(s)]^{-1}\, ds = \infty$. Hence $\theta(t) > \tau_P$ at every finite t and $\lim_{t\to\infty} \theta(t) = \tau_P$; the trajectory never reaches the P tie. Initialization at or below it instead collapses. Global interior collapse therefore holds above a_5, not at equality.
- At a_6, all interior skills select P. For every AI quality, the exact endpoints remain invariant: $\theta(t; 0) = 0$ and $\theta(t; 1) = 1$.

These critical values belong to this specification, not to the general theory. The next subsection holds the learning drifts fixed and examines how workflow incentives change the long-run outcome.

D.2 Details from Figure 1c: limit shifts by workflow incentives

We now vary only cost sensitivity γ, holding all other primitives and hence all three learning drifts fixed. The question is whether the same verified learning equilibrium remains active, or whether manual work supplies a different source of skill support.

Use the baseline primitives from Appendix D.1.1, fixing $\theta_a = 1/2$ and replacing the baseline self-checking cost and repair effort by $c_s(\theta) = 0.04 + 0.05(1 - \theta)$ and $\lambda = 7/10$. Both are fixed across all cases below. Following Figure 1c, Cases I–III are ordered by decreasing cost sensitivity, $\gamma = 3/4, 1/3, 1/4$, respectively. The shared primitives are

$$r_m(\theta) = r_v(\theta) = 2\theta - \theta^2, \qquad \varepsilon_a = \bar{\ell}_a = \tfrac{1}{2}, \qquad (\kappa, \Delta, k_a, \lambda) = (\tfrac{5}{4}, 3, 0, \tfrac{7}{10}),$$
$$c_g(\theta) = c_v(\theta) = \tfrac{1}{2} - \tfrac{2}{5}\theta, \qquad c_s(\theta) = \tfrac{9}{100} - \tfrac{1}{20}\theta.$$

In particular, $\ell_M(\theta) = (1 - \theta)^2$, $q(\theta) = \theta - \theta^2/2$, and $U_P = -1/4$. Since $c_g = c_v$ and $k_a = 0$, the utility gap is

$$D(\theta) = U_V(\theta) - U_M(\theta) = \underbrace{\ell_M(\theta) - \ell_V(\theta)}_{\text{output-quality gain}} + \gamma \underbrace{[c_s(\theta) - \tfrac{7}{10} q(\theta) c_g(\theta)]}_{\text{self-checking cost minus expected repair cost}}.$$

Both the output-quality gain and the unweighted cost difference are fixed. Increasing γ favors manual work relative to verified delegation wherever $c_s < (7/10) q c_g$, because expected repair is then more costly than manual self-checking. At the same time, increasing γ makes costless pure delegation more attractive relative to both workflows. These incentive changes leave $f_M = (13/4)\theta(1 - \theta)^2$, $f_P = -3\theta^2(1 - \theta)$, and

$$f_V(\theta; 1/2) = \theta(1 - \theta)\left[\frac{5}{4} - \frac{9}{4}\theta - \frac{1}{2}\theta^3\right]$$

unchanged. The verified bracket decreases strictly from $5/4$ to $-3/2$, so every case has the same stable branch root $\theta_V^\dagger \approx 0.524$, with $f_V > 0$ below it and $f_V < 0$ above it.

Proposition D.1 (Common learning dynamics, different long-run regimes)**.** *Under these common primitives, the following values of γ generate three different limits from the same initial skill $\theta_0 = 0.99$:*

Case	γ	*Optimal workflows as skill increases*	$\theta^\infty(0.99)$
I	$3/4$	$P \Rightarrow V \Rightarrow M \Rightarrow V$	$\tau_M^+ \approx 0.908$
II	$1/3$	$P \Rightarrow V$	0
III	$1/4$	$P \Rightarrow V$	$\theta_V^\dagger \approx 0.524$

More generally, for $\theta_0 \in (0,1)$:

1. *In Case I, $\tau_P \approx 0.742$, $\tau_M^- \approx 0.871$, and $\tau_M^+ \approx 0.908$, where $\tau_M^- := \tau^{(1)}$ and $\tau_M^+ := \tau^{(2)}$. Under the vanishing-step convention of Appendix C.2,*

$$\theta^\infty(\theta_0) = \begin{cases} 0, & 0 < \theta_0 < \tau_M^-, \\ \tau_M^+, & \tau_M^- < \theta_0 < 1. \end{cases}$$

   At $\theta_0 = \tau_M^-$, selecting M yields τ_M^+ and selecting V yields 0.

2. *In Case II, manual work is never optimal and $\tau_P \approx 0.551 > \theta_V^\dagger$. Every interior initial condition converges to 0.*

3. *In Case III, manual work is never optimal and $\tau_P \approx 0.502 < \theta_V^\dagger$. The limit is 0 for $\theta_0 < \tau_P$ and $\theta_V^\dagger$ for $\theta_0 > \tau_P$. At $\theta_0 = \tau_P$, selecting P yields 0 and selecting V yields $\theta_V^\dagger$.*

The priority $P \succ V \succ M$ used in Appendix D.1 selects 0 at both repelling ties described above.

*(*Local persistence.*) For each listed value $\gamma_0 \in \{3/4, 1/3, 1/4\}$, there is an open interval $J \subset (0,\infty)$ containing γ_0 on which the workflow ordering and the attractor type reached from $\theta_0 = 0.99$ are unchanged, with all other primitives and the tie rule fixed. On the Case I interval the limit is the continuously varying boundary $\tau_M^+(\gamma)$; on the Case II and III intervals it remains 0 and the same active root $\theta_V^\dagger$, respectively.*

Proof: Using $H_{MP} = U_M - U_P$ and $H_{VP} = U_V - U_P$, substitution gives $H_{MP}(\theta) = -\theta^2 + 2\theta - 3/4 + \gamma(9\theta/20 - 59/100)$ and

$$H_{VP}(\theta) = \frac{1}{2}\theta^4 - 2\theta^3 + \frac{9}{4}\theta^2 - \frac{1}{2}\theta + \gamma\left[-\frac{7}{50}\theta^3 + \frac{91}{200}\theta^2 + \frac{1}{20}\theta - \frac{1}{2}\right].$$

The roots that determine the optimal workflow are

Case	$H_{VP} = 0$	$D = H_{VP} - H_{MP} = 0$
I	0.742	$0.871,\ 0.908$
II	0.551	$\varnothing$
III	0.502	$\varnothing$

These counts are exact: if $v_p(x)$ denotes the sign variations in the Sturm sequence of p at x, then $(v_p(0), v_p(1)) = (2,1)$ for both H_{VP} and H_{MP} in all three cases. For D, the pairs are $(3,1)$, $(2,2)$, and $(2,2)$, respectively. Every sequence ends in a nonzero constant, so the roots are simple. The omitted H_{MP} roots are 0.752, 0.618, and 0.589, respectively, each above the corresponding H_{VP} root. Since both gaps are negative at 0, pure delegation is optimal below τ_P, the unique H_{VP} root.

Above τ_P, D determines the choice between M and V. In Case I, $D > 0$ on $(\tau_P, \tau_M^-) \cup (\tau_M^+, 1)$ and $D < 0$ on (τ_M^-, τ_M^+). In Cases II and III, D has no interior root and $D(0) = 39/50$ and $309/400$, respectively, so $D > 0$ throughout. This proves the workflow orderings.

The phase-line signs give the limits. In Case II, $\theta_V^\dagger < \tau_P$ makes $f_V < 0$ throughout R_V. Moreover, $f_V(\tau_P) < 0$, so every trajectory starting above τ_P crosses into R_P in finite time and then converges

to 0. Both tied branches have negative drift at τ_P. In Case III, $\tau_P < \theta_V^\dagger$ places the common root inside R_V, where f_V points toward it on both sides. Below τ_P, $f_P < 0$. At the repelling boundary, $f_P(\tau_P) < 0 < f_V(\tau_P)$, so the first update enters the basin selected by the tie rule.

In Case I, $\theta_V^\dagger < \tau_P < \tau_M^- < \tau_M^+$, so the verified root is again inactive. However, $f_M(\tau_M^+) > 0 > f_V(\tau_M^+)$ makes $\tau_M^+ \in S_{MV}$ under the vanishing-step convention. All trajectories above τ_M^- approach this boundary, whereas all active drifts below τ_M^- are negative. The lower boundary is repelling: $f_V(\tau_M^-) < 0 < f_M(\tau_M^-)$. Its first update therefore selects the lower basin under V and the upper basin under M, completing the classification.

Local persistence in γ. Fix one of the listed values γ_0. For each of $p \in \{H_{MP}, H_{VP}, D\}$, write $p(\theta, \gamma)$ to display its parameter dependence. Its coefficients vary smoothly with γ. At γ_0, every interior root is simple and both endpoint values are nonzero. The implicit-function theorem continues each root as a continuously differentiable function of γ. Choose disjoint neighborhoods of these roots with closures in $(0, 1)$ on which $|\partial_\theta p(\cdot, \gamma_0)|$ is bounded away from zero. On their compact complement in $[0, 1]$, $|p(\cdot, \gamma_0)|$ also has a positive lower bound. These bounds persist for sufficiently small $|\gamma - \gamma_0|$, excluding additional roots and preserving the signs between the continued roots. If p has no root, the latter bound applies on all of $[0, 1]$. The strict ordering of the pairwise utility roots is preserved after shrinking the parameter interval, so the workflow ordering is unchanged.

The learning drifts and $\theta_V^\dagger$ are independent of γ. At the three reference values, the strict inequalities are

$$\begin{aligned} &\text{I}: && \theta_V^\dagger < \tau_P < \tau_M^- < \tau_M^+ < 0.99, \\ &\text{II}: && \theta_V^\dagger < \tau_P < 0.99, \\ &\text{III}: && \tau_P < \theta_V^\dagger < 0.99. \end{aligned}$$

Continuity preserves these inequalities on a smaller open interval $J \subset (0, \infty)$ about γ_0. The active drift signs and the attracting manual–verified boundary in Case I therefore persist, and $\theta_0 = 0.99$ remains in the corresponding basin. The preceding vanishing-step phase-line argument gives the claimed limits throughout J. □

Figure 1c separates the two support mechanisms. Cases II and III share the ordering $P \Rightarrow V$, but τ_P lies on opposite sides of $\theta_V^\dagger$: verification sustains its own equilibrium only in Case III. Case I shows that an inactive verified root need not imply collapse, because a manual-work interval creates an alternative switching-supported limit. Starting from $\theta_0 = 0.99$, workers in all three cases initially choose V and follow exactly the same drift. Their paths diverge only after reaching different workflow boundaries. Thus identical initial behavior and learning opportunities need not predict identical retained skill; the incentives encountered as skill changes determine which support mechanism remains available.

The comparison also shows that long-run skill need not decrease with cost sensitivity. Ordered by increasing γ, Cases III, II, and I yield $\theta^\infty(0.99) \approx 0.524, 0, 0.908$, respectively. The pure-delegation region expands with γ, but at $\gamma = 3/4$ manual work becomes optimal on $(\tau_M^-, \tau_M^+) \approx (0.871, 0.908)$, creating a separate high-skill basin. For example, at $\theta = 0.9$, $c_s = 0.045 < 0.7qc_g = 4.851 \times 10^{-2}$ and $D = (2.55 - 3.51\gamma) \times 10^{-3}$; thus $D < 0$ at $\gamma = 3/4$, while $H_{MP} = 81/800 > 0$. Manual work is then cheaper than verification and strictly preferred to both alternatives. Workers starting above τ_M^- retain skill through this manual-work region, even though workers below it collapse and the common verified root remains inactive.

Unlike the menu comparison in Appendix D.4, V remains available throughout this construction. Incentives determine whether its stable root is active or manual work provides an alternative source of support.

D.3 Comparing dynamics with and without verification

Fix AI quality and all primitives, and compare the available workflows $\mathcal{A}_0 = \{M, P\}$ and $\mathcal{A}_1 = \{M, V, P\}$. Writing $W_i(\theta) := \max_{A \in \mathcal{A}_i} U_A(\theta)$, feasibility alone gives

$$W_1(\theta) \geq W_0(\theta) \qquad \text{for every } \theta.$$

Thus adding verified delegation can never reduce optimized *current* utility at a fixed skill level. It need not, however, improve the induced skill dynamics: utility determines which workflow is selected, whereas the selected workflow determines skill drift.

The branch ordering from Appendix C.4,

$$f_M(\theta) \geq f_V(\theta) \geq f_P(\theta), \qquad \theta \in (0,1),$$

makes this distinction explicit. If adding V replaces pure delegation P, the local skill drift weakly increases. If instead V replaces manual work M, the local drift weakly decreases. Hence the dynamic effect of verification depends on which workflow it displaces.

The classifications below use Assumption C.3 for each menu. In particular, manual drift is strictly positive on $(0,1)$.

Without verification, the interior threshold repels. Under $\mathcal{A}_0$, U_P is independent of worker skill and U_M is nondecreasing, so strict P choices lie below strict M choices; either region may be absent. Since

$$f_P(\theta) = -\Delta\theta^2(1-\theta) < 0, \qquad f_M(\theta) > 0$$

in the interior, a trajectory starting in a strict P region moves toward 0, while one starting in a strict M region moves toward 1. At a direct interior P/M switch, the two drifts point away from the boundary. Thus the M/P benchmark has no positive interior attractor, although it may have a high-skill basin converging to the endpoint 1.

Verification creates new mechanisms for sustaining positive skill. Adding V makes possible two types of positive interior attractor that are absent from the M/P benchmark. First, a stable verified-delegation root $z \in S_V$ is an active equilibrium: verification practice and error-triggered repair balance non-use decay inside the region where V is selected. Second, an M-to-V switching boundary can itself be attracting,

$$\tau \in S_{MV}, \qquad f_M(\tau) > 0 > f_V(\tau),$$

so that manual practice below the boundary offsets skill loss under verification above it. Appendix C.2 shows that these exhaust the positive interior attractors in the finite regular case.

Critical boundary roots require separate treatment: when $f_V(\tau_P) = 0$, a trajectory may approach τ_P from R_V without entering P, as at a_5 in Appendix D.1.2. This is not covered by the finite-regular attractor classification.

These additional sustaining mechanisms do not imply that introducing verification improves long-run skill from a common initial state. Appendix D.4 next holds all primitives fixed and changes only the available menu: adding V can strictly increase current optimized utility yet displace skill-building manual work and eliminate an otherwise high-skill basin.

Even when V is available, its learning branch alone does not determine whether competence is sustained. A stable verified root matters only if workflow incentives keep it active, while a suitably placed manual region can instead support an attracting switching boundary. The preceding comparison in Appendix D.2 holds f_M, f_V, f_P fixed and varies only workflow incentives, showing how the same learning dynamics can generate an active verified equilibrium, boundary-supported competence, or global collapse. These two comparisons separate the effect of expanding the workflow menu from the effect of changing incentives within a fixed menu.

D.4 Short-term utility gains, long-term skill loss

We hold all primitives fixed and vary only the available workflows, $\mathcal{A}_0 = \{M, P\}$ versus $\mathcal{A}_1 = \{M, V, P\}$. Use the specification in Appendix D.1.1, except that $\gamma = 0.32$, and initially fix $\theta_a = 1/2$. For $i \in \{0,1\}$, let $\theta^{\infty}_{\mathcal{A}_i}(\theta_0)$ denote the limiting skill under current-utility maximization over $\mathcal{A}_i$, using the vanishing-step convention of Appendix C.2.

Proposition D.2 (A statically beneficial option can eliminate the high-skill basin)**.** *Let* $\tau_{MP} = 0.62$ *and define*

$$W_0(\theta) = \max\{U_M(\theta), U_P\}, \qquad W_1(\theta) = \max\{U_M(\theta), U_V(\theta), U_P\}.$$

Then $W_1(\theta) \geq W_0(\theta)$ on $[0,1]$, with strict inequality on $(\tau_{MP}, 1)$. Nevertheless, every $\theta_0 \in (\tau_{MP}, 1)$ satisfies

$$\theta^{\infty}_{\mathcal{A}_0}(\theta_0) = 1, \qquad \theta^{\infty}_{\mathcal{A}_1}(\theta_0) = 0,$$
$$W_1\big(\theta^{\infty}_{\mathcal{A}_1}(\theta_0)\big) - W_0\big(\theta^{\infty}_{\mathcal{A}_0}(\theta_0)\big) = -\frac{513}{2500} < 0.$$

The last comparison concerns per-task utility at limiting skill, not cumulative or discounted lifetime utility.

Proof: *Workflow partitions.* Under $\mathcal{A}_0$, substitution gives $U_M(\theta) - U_P = -(\theta - 1.54)(\theta - 0.62)$. Thus P is strictly optimal below $\tau_{MP} = 0.62$, and M above it. Under $\mathcal{A}_1$,

$$U_V(\theta) - U_M(\theta) = \tfrac{3}{625} + 0.258(1-\theta)^2 + 0.032(1-\theta)^3 + 0.5(1-\theta)^4 > 0 \quad \text{on } [0,1],$$

so M is never selected. The remaining comparison is

$$H_{VP}(\theta) := U_V(\theta) - U_P = 0.5\theta^4 - 2.032\theta^3 + 2.354\theta^2 - 0.452\theta - 0.16.$$

Since $H'''_{VP}(\theta) = 12\theta - 12.192 < 0$, the derivative H'_{VP} is strictly concave. Together with $H'_{VP}(0) = -0.452 < 0 < H'_{VP}(1) = 0.16$, this gives a unique derivative zero in $(0,1)$. Hence H_{VP} first decreases and then increases; $H_{VP}(0) = -0.16 < 0 < H_{VP}(1) = 0.21$ implies a unique utility crossing $\tau_{VP} \approx 0.532$. In increasing order of skill, the partitions are $P \Rightarrow M$ under $\mathcal{A}_0$ and $P \Rightarrow V$ under $\mathcal{A}_1$.

Current utility. Menu inclusion gives $W_1 \geq W_0$ everywhere. For $\theta \in (\tau_{MP}, 1)$, the partitions and $U_V > U_M$ give $W_1(\theta) = U_V(\theta) > U_M(\theta) = W_0(\theta)$.

Skill and terminal utility. The common verified drift is

$$f_V(\theta) = -\frac{1}{4}\theta(1-\theta)(2\theta^3 + 9\theta - 5).$$

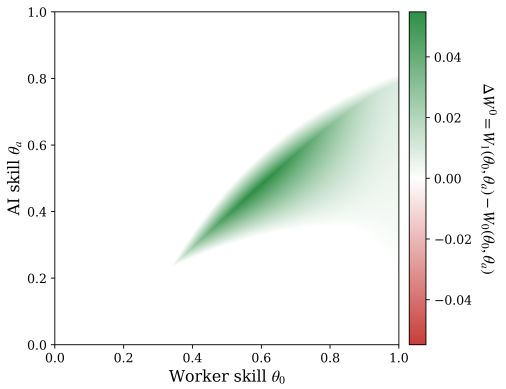
The cubic is strictly increasing, with derivative $6\theta^2 + 9 > 0$, and has a unique interior root $\theta^{\dagger}_V \approx 0.524$. At $\theta = 0.53$, the cubic is approximately $0.068 > 0$, whereas $H_{VP}(0.53) \approx -1.387 \times 10^{-3} < 0$; hence $\theta^{\dagger}_V < 0.53 < \tau_{VP}$. Thus $f_V > 0$ below $\theta^{\dagger}_V$ and $f_V < 0$ above it: the isolated verified equilibrium is stable but inactive under the optimal menu. For any $\theta_0 \in (\tau_{MP}, 1)$, the $\mathcal{A}_0$ trajectory remains in M, where $f_M > 0$, and therefore converges to 1. The $\mathcal{A}_1$ trajectory instead decreases through the active V-region. Since $f_V(\tau_{VP}) < 0$ and $f_P(\tau_{VP}) < 0$, it crosses the boundary into P and converges to 0. Finally, $W_0(1) = -28/625$ and $W_1(0) = -1/4$, giving $W_1(0) - W_0(1) = -513/2500$. □

Current utility and utility at limiting skill. We now vary (θ_0, θ_a), retaining $\gamma = 0.32$ and all other primitives above. Restore the AI-quality argument in $W_i(\theta; \theta_a) := \max_{A \in \mathcal{A}_i} U_A(\theta; \theta_a)$, and define $W^{\infty}_i(\theta_0, \theta_a) := W_i(\theta^{\infty}_{\mathcal{A}_i}(\theta_0; \theta_a); \theta_a)$. The differences at the common initial skill and at each system's own limiting skill are

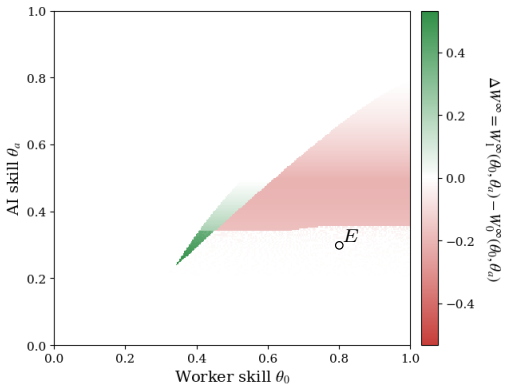
$$\Delta W^0(\theta_0, \theta_a) := W_1(\theta_0; \theta_a) - W_0(\theta_0; \theta_a),$$
$$\Delta W^{\infty}(\theta_0, \theta_a) := W^{\infty}_1(\theta_0, \theta_a) - W^{\infty}_0(\theta_0, \theta_a).$$

Figure 4a shows $\Delta W^0 \geq 0$: the worker can ignore the additional option. Its green region identifies strict current gains. Yet Figure 4c shows that, at $\theta_a = 1/2$, every $\theta_0 \in (0.62, 1)$ trades full skill for collapse. Adding V removes skill-building manual work without making its own stable learning state reachable: $\theta^{\dagger}_V$ lies below τ_{VP}, so the worker enters pure delegation before reaching it. Neither manual support nor an active verified equilibrium survives. Current utility therefore rises while the future capability supporting that utility disappears. At a fixed stable verified root z, $U_V(z) > U_P(z)$ is equivalent to $\gamma < \gamma_{\text{crit}}(z) := \frac{q(z)[\bar{\ell}_a - \ell_M(z)]}{c_v(z) + \lambda q(z) c_g(z)}$, provided the denominator is positive. This threshold determines whether verification beats pure delegation at z; the root is active only if verification also beats manual work.

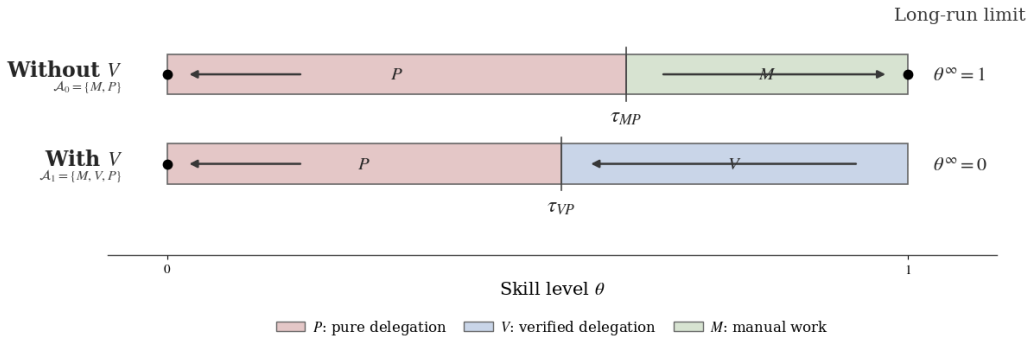
Figure 4b also shows losses without complete collapse or an initial strict gain. At the marked point $E = (\theta_0, \theta_a) = (0.8, 0.3)$, both menus initially select M: $U_M \approx -0.117$, $U_V \approx -0.123$, and



a: Current-utility gain from adding V.



b: Long-run per-task utility difference; $E = (0.8, 0.3)$.



c: Phase lines at $\theta_a = 1/2$; right-hand limits use the same $\theta_0 \in (0.62, 1)$.

Figure 4: **Adding verification can improve current utility while undermining skill sustainability.** Panels a and b compare current and limiting per-task utilities under $\{M, P\}$ and $\{M, V, P\}$, with $\gamma = 0.32$. Green/white/red indicate positive/zero/negative differences, with different color scales. Panel c shows how V displaces M without activating its own stable root. Critical initial states are excluded; plot-edge colors represent one-sided interior limits.

$U_P = -0.49$, so $U_M > U_V > U_P$ and $\Delta W^0 = 0$. Without V, skill rises to 1. With V, it reaches the upper M/V boundary $\tau_M^+ \approx 0.934$, where $f_M(\tau_M^+) > 0 > f_V(\tau_M^+)$ pins skill below full competence. The limiting utilities are approximately -0.045 without V and -0.060 with V, giving $\Delta W^\infty \approx -0.015$. Here verification limits further skill accumulation rather than causing complete decline: the cost emerges only after the initially identical manual trajectories encounter different future workflow options.

These losses are not uniform: the green region of Figure 4b also contains long-run gains. The comparison shows why current option value alone cannot establish long-run benefit; the selected workflows must also sustain the skill at which future utility is evaluated. The comparison in Appendix D.2 instead holds the full menu and learning drifts fixed while varying workflow incentives, separating the availability of an option from its activation.

E Details from Section 4: Learning Robustness and Finite-Task Dynamics

Appendix D changed workflow incentives or available options. Here we return to its unmodified baseline and hold current utilities fixed while varying learning and decay, then add learning from correct reviewed outputs. These exercises classify the deterministic conditional-mean dynamics. We subsequently retain a nonzero per-task increment and study stochastic first passage, which asks

a different question from the convergence theorems. The same subsection also reports increment sensitivity, Monte Carlo uncertainty, and the reproduction protocol. All reported simulation frequencies are illustrative computational results, not empirical estimates of worker behavior.

E.1 Effects of learning and decay rates

We hold current workflow incentives fixed and vary the learning mechanisms. All three exercises retain the functions and baseline parameters of Figure 1a–b, specified in Appendix D, except for the learning rates and additional learning channel varied below. The sweeps use 61 equally spaced values of $\kappa \in [0, 3]$ and 81 of $\Delta \in [0.1, 5]$, including endpoints. Neither parameter enters utility, so the workflow partition remains fixed within each exercise. Here $\ell_M(\theta) = (1-\theta)^2$ and $q(\theta) = \varepsilon_a(2\theta - \theta^2)$, suppressing the fixed θ_a argument. Write the verified drift as $f_V(\theta) = \theta(1-\theta)H(\theta)$, where

$$H(\theta) = (2q(\theta) + \kappa)(1-\theta) - \Delta(1-q(\theta))\theta.$$

The following lemma supplies the root-uniqueness condition used in both sweeps; no global assumption on arbitrary verification technologies is needed.

Lemma E.1 (A unique verified root in the two learning sweeps)**.** *Let $\varepsilon_a \in \{0.5, 0.7\}$, $q(\theta) = \varepsilon_a\theta(2-\theta)$, $\kappa \geq 0$, and $\Delta > 0$. The ratio*

$$\mathcal{R}(\theta) := \frac{(2q(\theta) + \kappa)(1-\theta)}{\theta(1-q(\theta))}$$

is strictly decreasing on $(0, 1)$. Consequently, any interior verified root is unique and attracting within its branch.

Proof: Write $e = \varepsilon_a$ and $Q(\theta) = 1 - e\theta(2-\theta) > 0$. Split the ratio into

$$\mathcal{R}(\theta) = \frac{2e(2-\theta)(1-\theta)}{Q(\theta)} + \kappa\frac{1-\theta}{\theta Q(\theta)}.$$

The derivatives of its two components, before multiplication of the second by κ, are

$$\frac{2e[e(2-\theta)^2 + 2\theta - 3]}{Q(\theta)^2}, \qquad \frac{e(2\theta^3 - 5\theta^2 + 4\theta) - 1}{\theta^2 Q(\theta)^2}.$$

The first numerator is negative: the bracket is convex in θ, and its endpoint values are $4e - 3 < 0$ and $e - 1 < 0$ for both stated error rates. For the second, $2\theta^3 - 5\theta^2 + 4\theta \leq 28/27$ on $[0, 1]$, with its maximum at $\theta = 2/3$. Since $e \leq 7/10$, its numerator is also strictly negative. Hence $\mathcal{R}' < 0$, including when $\kappa = 0$. Finally,

$$H(\theta) = \theta(1-q(\theta))(\mathcal{R}(\theta) - \Delta).$$

At a root, $H' = \theta(1-q)\mathcal{R}' < 0$, and therefore $f_V' = \theta(1-\theta)H' < 0$. Strict monotonicity permits at most one root. □

A branch root is sustainable only if it belongs to the strict verified region R_V and is reachable from the specified initial skill. The two subsections below check this compatibility at the fixed workflow boundaries.

Sustainable verification versus collapse. At $\theta_a = 0.5$, the workflow ordering is $P \to V$, with pure-delegation cutoff $\tau_P \simeq 0.521$. In this configuration, $\tau_P = \tau_{VP}$ in the pairwise notation of Appendix D.1.1; throughout this appendix, we use τ_P for the actual pure-delegation cutoff. We initialize skill at $\theta_0 = 0.8$, inside the verified region. The sign of $H(\tau_P)$ determines whether this trajectory reaches an active equilibrium or crosses into pure delegation. The critical value solves $H(\tau_P) = 0$:

$$\kappa_{\text{crit}}(\Delta) = \tfrac{(1-q(\tau_P))\Delta\tau_P}{1-\tau_P} - 2q(\tau_P) \simeq 0.669\Delta - 0.771.$$

If $\kappa > \kappa_{\text{crit}}(\Delta)$, the unique attracting root lies above τ_P, and the trajectory remains in V and converges to it. If $\kappa < \kappa_{\text{crit}}(\Delta)$, verified drift is negative on $[\tau_P, 1)$, so skill crosses into P and then converges to zero. At equality, skill approaches τ_P from above without crossing it in finite continuous time; this is a critical boundary limit, not collapse or an interior verified equilibrium. Exact initialization at a utility tie remains subject to the fixed tie rule in (1).

At the baseline, $\kappa_{\text{crit}}(3) \simeq 1.236$ and $\theta_V^\dagger \simeq 0.524$, leaving a margin of only 2.5×10^{-3} above τ_P. Figure 2a compares collapse and sustainable verified delegation across the stated grid. The linear boundary shows how verification practice offsets non-use decay while current workflow incentives remain unchanged.

Verified equilibria versus boundary-supported skill. At $\theta_a = 0.3$, the workflow ordering is $P \to V \to M \to V$, with boundaries approximately 0.370, 0.498, and $\tau_M^+ \simeq 0.935$. We initialize at $\theta_0 = 0.95$, in the upper verified region. Manual drift $f_M(\theta) = (2+\kappa)\theta(1-\theta)^2$ is positive on the manual interval immediately below τ_M^+. When $H(\tau_M^+) < 0$, verified drift is negative on $(\tau_M^+, 1)$, so the two branches point inward and support the limit $\theta^\infty = \tau_M^+$ under vanishing-step selection for the conditional-mean recursion (Appendix C.2). The threshold between this regime and an interior verified equilibrium is

$$\kappa_{\text{crit}}^{\text{upper}}(\Delta) = \frac{(1-q(\tau_M^+))\Delta\tau_M^+}{1-\tau_M^+} - 2q(\tau_M^+) \simeq 4.388\Delta - 1.394.$$

For $\kappa < \kappa_{\text{crit}}^{\text{upper}}(\Delta)$, competence is boundary-supported. For $\kappa > \kappa_{\text{crit}}^{\text{upper}}(\Delta)$, the unique attracting verified root lies in $(\tau_M^+, 1)$, and the trajectory converges to it while remaining in V. At equality, skill approaches τ_M^+ from the verified side; the limiting skill is again the boundary, but the verified drift vanishes there instead of pointing strictly inward.

At $(\kappa, \Delta) = (1.25, 3)$, $f_V(\tau_M^+) \simeq -0.041$, so the limit is $\tau_M^+ \simeq 0.935$. Figure 2b shows how the two positive-skill regimes divide the specified learning-parameter grid.

The two sweeps change skill formation without moving the workflow partition. They therefore test persistence of the deterministic regimes under learning parameter changes, not robustness to arbitrary changes in workflow costs or AI-error technology. We next modify the learning mechanism itself.

E.2 Additional learning from reviewing correct AI outputs

The baseline already includes verification practice of strength κ on every reviewed task. We now add learning from studying a correct AI output, contributing $\xi(1-\theta)$, $\xi \geq 0$, when the output is correct and reviewed. This additional channel changes only the verified branch:

$$f_V^{(\xi)}(\theta) = f_V(\theta) + \theta(1-\theta)^2(1-\varepsilon_a)\xi.$$

At fixed AI quality, it is equivalent within this branch to replacing κ by $\kappa_{\text{eff}} = \kappa + (1-\varepsilon_a)\xi$; manual learning, pure-delegation decay, and current workflow utilities remain unchanged.

We retain $\theta_a = 0.5$, $\theta_0 = 0.8$, and the grid of Appendix E.1. The critical value becomes

$$\kappa_{\text{crit}}(\Delta, \xi) = \kappa_{\text{crit}}(\Delta) - (1-\varepsilon_a)\xi = \kappa_{\text{crit}}(\Delta) - 0.5\xi.$$

The same strict-inequality and critical-equality classifications apply as in that exercise. Table 2 reports the active root at $(\kappa, \Delta) = (1.25, 3)$ and the fraction of grid points satisfying $\kappa < \kappa_{\text{crit}}(\Delta, \xi)$. These fractions describe the equally weighted parameter grid, not a population probability of skill collapse.

Table 2: **Additional learning from correct reviewed outputs.** Roots and margins use $(\kappa, \Delta) = (1.25, 3)$; collapse shares use the 61×81 learning-parameter grid.

ξ	$\kappa_{\text{crit}}(3, \xi)$	$\theta_V^\dagger$	$(\theta_V^\dagger - \tau_P)/10^{-3}$	Collapse share
0.00	1.236	0.524	2.5	0.339
0.10	1.186	0.532	11.3	0.326
0.25	1.111	0.545	23.7	0.307
0.50	0.986	0.564	43.0	0.277
1.00	0.736	0.598	76.4	0.222

Across the reported values, additional learning raises the verified equilibrium and widens its margin above the delegation boundary, while a positive collapse share remains. Because the frequency of

the added signal varies with AI error frequency, this fixed-quality exercise does not establish AI-quality comparative statics for the extension. The competing effects can nevertheless be separated within the verified branch. Holding ξ and the other primitives fixed, let $\Lambda(\theta) = \Delta\theta - \ell'_M(\theta) > 0$; differentiation gives

$$\partial_{\theta_a} f_V^{(\xi)} = \theta(1-\theta)\varepsilon'_a(\theta_a)\big[r_v(\theta)\Lambda(\theta) - \xi(1-\theta)\big].$$

Since $\varepsilon'_a \le 0$, better AI weakly reduces the verified drift where $r_v\Lambda \ge \xi(1-\theta)$; where the inequality reverses, learning from more frequent correct outputs can instead offset the reduction in repair opportunities. This is a pointwise branch comparison, not an ordering of the resulting workflow-dependent paths or long-run limits.

This extension preserves the distinction between a fixed-workflow effect and a selected-path comparison. We now return to $\xi = 0$ and ask whether a stable baseline mean equilibrium survives a finite sequence of randomly occurring repair opportunities.

E.3 Stochastic escape into pure delegation

We return to the baseline learning law ($\xi = 0$). The ODE describes conditional mean drift, whereas repair opportunities occur randomly over individual tasks. We examine first passage from verified into pure delegation, starting either near an active verified equilibrium or above it to include the convergence phase. These simulations concern finite-task durability of verified delegation; they do not classify stochastic long-run outcomes or test manual–verified boundary attractors.

Task-level updates and simulation configuration. We use the task-level update (6) and its potential-gradient signals from Appendix B.3. In this simulation subsection, t counts tasks, so N tasks correspond to exposure $N\eta$. Before task t, the worker selects $A_t = A^*(\theta_t;\theta_a)$ and subsequently recomputes the choice after the skill update. For histories currently in V, fresh error and detection draws give $R_t = E_t I_{D,t}$ with probability $q(\theta_t;\theta_a)$. Histories in M and P use their respective deterministic task signals.

The implementation applies projection onto $[0,1]$ as a safeguard. However, $L = 2$ in this baseline, and every increment used below satisfies (7). Projection is therefore inactive and the conditional mean increment is exactly $\eta f_{A_t}(\theta_t;\theta_a)$. The justification is the one-step calculation in Appendix B.3, not an identification of the expected random path with the ODE trajectory.

Setting	Configuration
Primitives	Baseline of Appendix D.1; $\theta_a = 0.5$, $\xi = 0$; only γ varies within each escape experiment.
Workflow tie priority	M, then V, then P. This retains the simulator's reported rule, distinct from the deterministic illustration's P, V, M priority. The difference is confined to exact ties.
Update timing	Select workflow before each task, draw events only for histories currently in V, update skill, then select again on the next task.
Histories	5,000 independent histories per parameter cell.
Equilibrium-initialized exercise	$\theta_0 = 523647/10^6$, $\eta = 5\times10^{-4}$, and $N \in \{5{,}000, 10{,}000, 20{,}000\}$.
Increment exercise	$\theta_0 = 0.8$, $N\eta = 10$, and $\eta \in \{10^{-3}, 5\times10^{-4}, 2.5\times10^{-4}\}$.
Escape statistic	First pre-task selection of P strictly before task index N; defined precisely below.

This configuration preserves the reported numerical experiments. The end of this subsection records the seed, draw order, and provenance limits.

Safety margins and escape probabilities. This finite-horizon illustration uses the specification in Figure 1a–b and Appendix D, fixes $\theta_a = 0.5$, and varies only γ among the baseline parameters. For the five values of γ in Table 3, the ordering remains $P \to V$. Varying γ shifts τ_P but leaves the verified conditional mean drift and its attracting root $\theta_V^\dagger \simeq 0.524$ unchanged. As in the main text, let $d = \theta_V^\dagger - \tau_P$ denote the safety margin. Define the first-passage time by

$$T_P := \inf\{t \in \mathbb{N}_0 : A^*(\theta_t;\theta_a) = P\}.$$

We use $\inf \varnothing := \infty$. Here P is selected exactly when $\theta_t < \tau_P$; the simulator selects V at the P/V tie.

All histories start at $\theta_0 = 523647/10^6 \simeq \theta_V^\dagger$, using a six-decimal representation of the verified equilibrium to avoid an appreciable initial convergence phase. Table 3 reports the fraction of 5,000 independent histories with $T_P < N$, using $\eta = 5 \times 10^{-4}$. The three task horizons correspond to $N\eta = 2.5, 5, 10$. Workflow choice is evaluated before each task, at $t = 0, \ldots, N-1$; entry only after the final update is excluded from this escape statistic and included in the separately recorded final P share.

Table 3: **Escape from an attracting conditional-mean equilibrium.** Entries estimate $\Pr(T_P < N)$ from 5,000 histories initialized at $523647/10^6 \simeq \theta_V^\dagger$, with $\eta = 5 \times 10^{-4}$.

γ	Margin $d/10^{-3}$	$N = 5{,}000$	$N = 10{,}000$	$N = 20{,}000$
0.280	13.85	0.067	0.140	0.285
0.285	10.98	0.177	0.327	0.545
0.290	8.14	0.376	0.601	0.823
0.295	5.31	0.646	0.836	0.958
0.300	2.50	0.882	0.963	0.998

In the reported simulations, escape is more frequent at smaller margins. Increasing the horizon enlarges the first-passage event; the table quantifies this accumulated risk. Although the verified conditional mean drift is locally restoring, entry into P changes the update to $\theta_{t+1} = \theta_t[1-\eta\Delta\theta_t(1-\theta_t)]$. At the stated step size, this is strictly between 0 and θ_t for interior skill, so the P region is absorbing. Thus local mean attraction does not guarantee survival in V over a finite task history. These probabilities characterize escape during the stated horizons, not a stochastic long-run classification; they do not contradict the conditional-mean attractor. We next separate sensitivity to the per-task increment from sampling uncertainty in the finite simulations.

Per-task increment sensitivity and Monte Carlo uncertainty. We next initialize at $\theta_0 = 0.8$ to include the initial convergence phase and vary η while holding time-scaled exposure at $N\eta = 10$, with all other parameters unchanged. The three increments 10^{-3}, 5×10^{-4}, and 2.5×10^{-4} therefore use $N = 10{,}000$, $20{,}000$, and $40{,}000$ tasks, respectively. Unlike a numerical solver tolerance, η scales each realized learning update; reducing it at fixed $N\eta$ increases the number of tasks and changes stochastic fluctuations. Each reported probability uses 5,000 histories and the same first-passage definition. Because initialization differs, these results are not directly comparable to the equilibrium-initialized probabilities in Table 3.

Table 4: **Sensitivity to the per-task increment.** Reported estimates of $\Pr(T_P < N)$ with $\theta_0 = 0.8$, $N\eta = 10$, and 5,000 histories per cell.

Margin $d/10^{-3}$	$\eta = 10^{-3}$	$\eta = 5 \times 10^{-4}$	$\eta = 2.5 \times 10^{-4}$
13.85	0.329	0.086	0.004
8.14	0.670	0.419	0.154
2.50	0.910	0.843	0.742

In these reported runs, smaller increments reduce escape probabilities, while narrower margins retain greater escape risk. The margin ordering persists over the tested range, but the probability levels vary substantially; this comparison establishes neither numerical convergence nor quantitative robustness of the equilibrium-initialized experiment. Assuming independent histories within a cell, a reported fraction $\widehat{p}$ has binomial Monte Carlo standard error $\sqrt{\widehat{p}(1-\widehat{p})/5000} \leq 7.1 \times 10^{-3}$, quantifying sampling uncertainty only. At $\eta = 5 \times 10^{-4}$, three separately seeded runs give probabilities $0.843, 0.849, 0.840$ for $d \simeq 2.5 \times 10^{-3}$, and $0.419, 0.408, 0.419$ for $d \simeq 8.14 \times 10^{-3}$. Each triple includes the corresponding central-column estimate, rather than three additional replications.

Reproduction protocol and provenance. The protocol below records the reconstructed recipe reported for the simulation tables. It distinguishes the mathematical experiment, specified above, from the provenance of the numerical outputs. The recipe is not an archived manifest of the original

runs; a version-pinned simulator, random number library, and saved output manifest are needed to preserve exact reproduction across environments.

For $\gamma_k = 0.24 + 0.005k$ (zero-based k), the reported reconstruction uses `default_rng` with seed $12345 + k$ and the six-decimal initialization, and reproduces all 15 displayed probabilities in Table 3. On each task, the simulator draws one vector of error uniforms and then one vector of detection uniforms for histories currently in V, in their fixed index order; repair requires both draws to succeed.

The same seed at each horizon makes the columns nested prefixes of the same histories, not independent replications across horizons.

The increment-sensitivity runs use $\theta_0 = 0.8$ and seed $12345 + k$. Replacing the base seed by 23456 and 34567 reproduces the other two runs, respectively. The reported reconstruction matches all nine increment-table probabilities and the four additional replication probabilities exactly; these seed checks quantify simulation variability, not empirical uncertainty.

These checks concern computational variability and reproduction of the reported simulations. They do not estimate empirical uncertainty about human learning rates, nor do they validate the magnitude of real-world escape risk.

F Details from Section 4: illustrative data-to-parameter mapping

We map performance and interaction records from Bastani et al. (2025) to selected model inputs and use them in a worked example. This data-to-parameter mapping neither jointly identifies the model nor tests its causal mechanism or long-run predictions. AI capability is fixed, and its argument is suppressed.

Scope, data, and workflow proxies. Some inputs come from sample summaries, conditional fits, or moment matches; others are assigned from proxies or fixed by normalization. Table 5 lists their sources and assumptions.

The analysis code is included in the supplementary material and will be publicly released with the paper. Reproduction also requires the underlying records and processing details listed below.

Table 5: Model inputs, construction, and interpretation in the illustrative mapping.

Model object	Construction and interpretation
Initial skill index θ_0	Conditional GPA–Control-score fit; index defined by the imposed reliability scale.
AI error rate ε_a	Auxiliary benchmark error counts; outputs and correctness labels are publicly available.
Error severity $\bar{\ell}_a$	Effective-loss fit to Base mean score at fixed ε_a.
Verification reliability r_v	Conditional Tutor-score moment match; detection and repair quality not separately observed.
Activity costs and k_a	Timing-based assignments; not event-adjusted structural estimates.
Cost weight γ	Time-tradeoff proxy; filtering details not documented here.
Non-use rate Δ	Conditional descriptive Base fit; not an independently measured forgetting rate.
Practice rate κ	Imposed 0.10; Control post-score moment remains unmatched.
Scale conventions	One episode-time unit, 20 minutes per cost unit, unit generative-learning scale, and $\lambda = 0.30$.
Illustrative trajectory	Worked example using the mapped inputs and specified update rule.

Experimental data and workflow proxies. For this mapping, Control proxies manual work M; GPT Base, which can supply complete solutions, proxies pure delegation P; and GPT Tutor, which supplies hints and guidance, proxies verified delegation V. These are arm-level proxies, not observed individual workflow choices: Tutor does not require reviewing a completed AI answer, and Base users may inspect or modify outputs.

The released repository supplies `final_data.csv` and timestamped chats. The reported sample is restricted to Year 9, Session 1, and excludes Honors students, the non-compliant class 9J, and students with missing previous GPA:

Condition	Students	Prior GPA	Part 2	Part 3
Control	113	0.871	0.442	0.416
GPT Base	72	0.859	0.528	0.330
GPT Tutor	98	0.876	0.620	0.456

Part 2 and Part 3 are mean normalized practice and immediate unaided scores. The four Part 2 problems form one episode-time unit: loss and event exposure are averaged per problem, whereas time and cost are summed across the block. Skill is held fixed when forming these exposures and updated afterward; q is not the probability of at least one repair in the episode. This aggregation approximates the data block without changing the one-task model. Process moments use their available observations, not a common complete-case sample.

Construction of model inputs. Standardize previous GPA across all 283 students using population variance:

$$z_i = (\text{gpa_prev}_i - \mu_g)/\sigma_g, \qquad \mu_g = 0.87, \quad \sigma_g = 0.077.$$

The released item records give four nonmissing binary scores per student, with mean `Part2Tot`. The reported Control fit uses

$$Y_i := 4\,\text{Part2Tot}_i, \qquad Y_i \mid z_i \sim \text{Binomial}(4, p_i), \qquad \text{logit}(p_i) = \alpha + \beta z_i.$$

The likelihood treats students as independent and the four responses as conditionally independent with common p_i, omitting item difficulty and residual dependence. The 113 Control students provide 200 correct answers out of 452. Reported maximum-likelihood coefficients are $\widehat{\alpha} = -0.267$ and $\widehat{\beta} = 0.538$; the constraint $\beta \geq 0$ does not bind.

Choosing the reliability scale $r_m(\theta) = 2\theta - \theta^2$ gives the GPA-based skill index

$$\widehat{p}_i = \text{logit}^{-1}(\widehat{\alpha} + \widehat{\beta} z_i), \qquad \theta_{0,i} = 1 - \sqrt{1 - \widehat{p}_i}.$$

The mean of the individual indices is approximately 0.26; inverting mean reliability is not equivalent. Applying this Control-practice mapping to all arms as initial reliability assumes transferability and the fixed-block approximation. The index is not independently measured pre-practice ability; subsequent calculations retain the displayed coefficients without rounding individual skills.

Error frequency and effective severity. The released benchmark file `gpt_answers_full.csv` contains ten GPT outputs per practice problem and response-level correctness labels. For the four practice problems used here, these labels give $8, 6, 0, 10$ incorrect outputs, yielding $\widehat{\varepsilon}_a = (8 + 6 + 0 + 10)/40 = 0.6$, a pooled per-problem error rate before student intervention.

Base students answer 152 of 72×4 items correctly, giving $\bar{s}_P = 19/36 \simeq 0.528$. Conditional on the P proxy and this error-rate input, matching $1 - \varepsilon_a \bar{\ell}_a$ gives

$$\widehat{\bar{\ell}}_a = (1 - \bar{s}_P)/0.60 = 85/108 \simeq 0.787.$$

Literal binary error loss is one; this fitted severity instead absorbs the mismatch between saved AI outputs, student responses, and the P proxy.

Verification reliability. Use $r_v(\theta; \rho_r) = 1 - (1 - \theta^{\rho_r})^2$, $\rho_r > 0$. Two independent per-pass detection opportunities with success probability θ^{ρ_r} motivate the form, but are not observed for every Tutor interaction. Under the workflow proxy, initial-skill mapping, and manual-repair convention,

$$\widehat{s}_{V,i}(\rho_r) = 1 - \varepsilon_a \left[(1 - r_v(\theta_{0,i}; \rho_r))\bar{\ell}_a + r_v(\theta_{0,i}; \rho_r)(1 - r_m(\theta_{0,i})) \right].$$

Conditional on these conventions, the reported moment match chooses ρ_r to reproduce the equally weighted Tutor mean,

$$\frac{1}{98} \sum_{i \in \text{Tutor}} \widehat{s}_{V,i}(\rho_r) = \frac{243}{392} \simeq 0.620,$$

yielding $\widehat{\rho}_r \simeq 0.748$, implied mean $r_v \simeq 0.589$, and an implied mean per-problem repair frequency $q \simeq 0.353$. This frequency is not a directly observed repair rate or a successful-correction probability. Subsequent calculations use

$$\varepsilon_a = 0.6, \quad \bar{\ell}_a = 0.79, \quad r_v(\theta) = 1 - (1 - \theta^{0.75})^2, \quad q(\theta) = 0.6 r_v(\theta).$$

Without linked error, detection, and repair outcomes, the fit does not identify detection separately from repair quality. Also, $\rho_r < 1$ makes r'_v diverge at zero: this mapping is smooth at positive skill but is not a C^1 specialization of the baseline on $[0, 1]$.

Timing proxies. Set $c_j(\theta) = a_j(2 - \theta)$, $j \in \{g, s, v\}$, with 20 episode-minutes per cost unit. Chats cover 72 Base and 94 Tutor students. Within each problem thread, order prompts by time, cap successive gaps at five minutes, and sum gaps across threads to obtain A_i^{time}. Let S_i be the first-to-last prompt span and P_i the prompt count; exclude records crossing dates or having $A_i^{\text{time}} > S_i$.

Cleaned Tutor means give $s \simeq 9.823/11.955 \simeq 0.822$ minutes per prompt. Pricing the Base mean of 8.087 prompts gives $k_a = s(8.087)/20 \simeq 0.33$. Define

$$y_{v,i} = \frac{s(P_i - 8.087)}{20}, \qquad y_{\text{gap},i} = \frac{S_i - A_i^{\text{time}}}{20}.$$

Reported means correspond to 3.178 and 4.457 minutes. Extra prompts proxy inspection; gaps may include repair, waiting, or unrecorded activity. Part 3 duration proxies manual cost for 249 students (101 Control, 64 Base, 84 Tutor), paired with the post-skill index defined below. Transferring it to Part 2 assumes comparable total effort despite possible differences in item count and difficulty. The final cleaned chat sample sizes and the treatment of negative extra-prompt values are not documented here.

Event weighting and coefficient assignments. With four additive comparable problem costs, c_g is episode generation cost and $\lambda c_g/4$ is conditional per-problem repair cost; expected episode repair cost is $q\lambda c_g$. If gaps measured only repair time,

$$\mathbb{E}[y_{\text{gap},i} \mid \theta_i, V] = q(\theta_i)\lambda a_g(2 - \theta_i).$$

Thus a regression on $2 - \theta_i$ alone cannot identify λa_g; it requires the regressor $q(\theta_i)(2 - \theta_i)$. The reported one-coefficient summaries are $b_{\text{gap}} \simeq 0.128$, $b_v \simeq 0.093$, and $b_M \simeq 0.465$. The illustration assigns $\lambda a_g = b_{\text{gap}}$, $a_v = b_v$, and $a_g + a_s = b_M$, with imposed $\lambda = 0.30$ fixing the otherwise undetermined generation/self-check split. The resulting rounded illustrative cost inputs are

$$c_g(\theta) = 0.43(2 - \theta), \quad c_s(\theta) = 0.04(2 - \theta), \quad c_v(\theta) = 0.09(2 - \theta), \quad k_a = 0.33.$$

These are assignments, not event-adjusted estimates; in particular, $c_g > c_v > c_s > 0$ is an imposed illustrative ordering.

Cost preference. Let t_i be quiz minutes offered for AI assistance and $\Delta\ell_i = \ell_M(\theta_{0,i}) - \ell_{\text{AI},i}$, with $\ell_{\text{AI},i} = \ell_P$ for Base and $\ell_V(\theta_{0,i})$ for Tutor. If surrendered quiz time and completion time have the same marginal value, and surrender has no separate score effect, indifference gives

$$\Delta\ell_i = \gamma_i \left[\text{cost}_{\text{AI},i} - \text{cost}_{M,i} + \frac{t_i}{20} \right].$$

The reported conversion instead treats t_i as a performance-only valuation net of completion-cost differences: $\Delta\ell_i = \gamma_i t_i/20$. Hence $\gamma_i = 20\Delta\ell_i/t_i$ requires $t_i > 0$ and, for a nonnegative weight, $\Delta\ell_i \geq 0$. The reported median of $\Delta\ell_i/t_i$ over finite observations is $9330728/10^9$, yielding $\gamma \simeq 0.187$, rounded to 0.19. This is a provisional proxy under the net-valuation assumption. The treatment of zero willingness to pay and negative gains in the reported median is not documented here.

Learning rates and identification limits. Map unaided scores to $\theta_{1,i}^{\text{obs}} = 1 - \sqrt{1 - \text{Part3Tot}_i}$. Unlike $\theta_{0,i}$, this inverts a noisy observed score rather than fitted reliability. A common pre/post skill scale does not separate score noise, problem difficulty, or within-practice learning. Normalizing episode duration and generative-learning scale to one gives the descriptive unit-step updates

$$\theta_{1M} = \theta_0 + (2 + \kappa)\theta_0(1 - \theta_0)^2, \qquad \theta_{1P} = \theta_0 - \Delta\theta_0^2(1 - \theta_0).$$

All 72 Base students have post-scores. With $x_i = \theta_{0,i}^2(1-\theta_{0,i})$, minimizing

$$\sum_{i\in\text{Base}} [\theta_{1,i}^{\text{obs}} - \theta_{0,i} + \Delta x_i]^2$$

gives

$$\widehat{\Delta} = \max\left\{0, \frac{\sum_{i\in\text{Base}} x_i(\theta_{0,i} - \theta_{1,i}^{\text{obs}})}{\sum_{i\in\text{Base}} x_i^2}\right\}.$$

The denominator is positive. The reported fit is 0.951; we use $\Delta = 0.95$. Re-estimation requires individual records and the GPA mapping, not sample means alone. The fit is conditional on score inversion and Base representing non-use; it is not an independently measured forgetting rate.

Before normalization, episode duration h and generative scale a enter these moments through $h(2a+\kappa)$ and $h\Delta$. Setting $h = 1$ fixes time units, whereas $a = 1$ is a separate restriction. For the worked example, we set $\kappa = 0.10$, five percent of the generative coefficient 2, rather than estimate it from Control outcomes.

The logistic intercept matches mean initial Control reliability to $200/452 \simeq 0.442$, up to rounding. With $a = 1$ and $\kappa = 0.10$, the manual update remains in $(0, 1)$ and strictly raises each predicted reliability, so mean predicted post-score exceeds 0.442, rather than matching the reported 0.416. This mismatch cannot be interpreted as a statistical rejection of learning: the common scale approximation leaves noise and task difficulty unresolved. One pre/post episode does not identify or test repeated-learning dynamics.

Conditional illustrative calculation. For Tutor student 130122 (class 9T), GPA is $907451/10^6$, Part 2 score 0.75, and Part 3 score 0.60. This selected in-sample case is a worked example, not a representative or held-out test. Using the displayed GPA normalization and fitted coefficients gives $z_i \simeq 0.486$ and $\theta_{0,i} \simeq 0.292$.

We use this index without intermediate rounding and the functions above, with $(\gamma, \lambda, \Delta, \kappa) = (0.19, 0.30, 0.95, 0.10)$ and unit episode and generative scales. Initially $r_v \simeq 0.637$, $q \simeq 0.382$, and $(U_M, U_V, U_P) \simeq (-0.654, -0.472, -0.537)$, so V is strictly optimal. At each episode, we recompute utilities and take a unit Euler step of the selected branch. Under V,

$$\theta_{t+1} = \theta_t + \theta_t(1-\theta_t)\left[(2q(\theta_t) + 0.10)(1-\theta_t) - 0.95(1-q(\theta_t))\theta_t\right].$$

The first update gives $\theta_{1,i} \simeq 0.383$ and a model-implied unaided score $r_m(\theta_{1,i}) \simeq 0.619$. Across 1,000 iterations of this recurrence, V remains optimal and skill approaches the active root $\theta_V^\dagger \simeq 0.759$; the manual–verified cutoff is about 0.924.

These values follow the specified discrete recurrence, not the exact continuous-time dynamics, and are conditional model outputs rather than forecasts of the student's future performance.

Interpretation and remaining measurements. The calculation shows how the available records can be translated into an illustrative model instance, but it does not jointly identify the underlying mechanisms. Doing so would require event-linked AI errors, detection and repair outcomes, activity-specific effort, repeated workflow choices, and comparable unaided skill measures across episodes.